Validity and Reliability Evidence for an Experimental Performance Evaluation Instrument for Educational Speech-Language Pathologists

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Jill Rentmeester Disher

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Mark DeRuiter, Ph.D. and Benjamin Munson, Ph.D.

February 2018

Acknowledgements

I would like to acknowledge my faculty advisors, Drs. Mark DeRuiter and Benjamin Munson, for their unwavering support and guidance throughout my doctoral studies and in particular, this doctoral dissertation. Your encouragement, wisdom, scholarly excellence, and patience have been deeply appreciated and I am forever grateful.

I would like to thank my dissertation committee members, Drs. Liza Finestack and Kristen McMaster, for their thoughtful feedback during the dissertation prospectus process and review of writing for the final thesis.

Special thanks to doctoral students Giang Pham, PhD, Kerry Danahy Ebert, PhD, Christine Wing, PhD, Bita Payesteh, PhD, and Hannah Julien for their support and inspiration during our doctoral studies and speech-language pathologists Cathy DeRuiter, Heidi Hueffmeier, Anna Lakin and Cindy Vachon for their feedback and friendship during this project.

Finally, I would like to thank my husband, Larry Disher; our beautiful daughter, Charli Disher; parents, Eric and Linda Rentmeester; and close friends and family. Your encouragement during the dissertation process did not go unnoticed!

Abstract

Performance evaluation for educators is intended to measure, develop, and support professional practices, and, in turn, improve student outcomes. To date, however, very little research exists to support the performance evaluation practices for non-classroom educators (Holdheide, Goe, Croft, & Reschly, 2010), such as educational speech-language pathologists (SLPs).

Validity and reliability evidence for an experimental performance evaluation instrument specifically designed for SLPs was examined in this study. Study data were from 111 SLPs in a mid-size urban district who were evaluated one during an academic school year. The performance of the 111 SLPs was also described, so that any potential bias in the instrument could be examined.

Results showed a restricted range of performance in which most SLPs were rated as *proficient* or *exemplary* on performance evaluation items. Some preliminary indications of bias were present, such that SLPs serving birth-five students, students with combined communication disorders, or students with moderate-severe disabilities were rated 6-10 points lower in total score (out of a maximum score of 108) compared to their colleagues who did not serve those populations. Construct validity analyses showed that the instrument's items were only loosely related to each other, although exploratory factor analyses did suggest an underlying structure of four domains. Face validity was gauged through optional perception surveys of the 111 SLPs in which the majority of survey participants felt the instrument items represented effective SLP practices. Finally, performance evaluation items demonstrating the highest evidence of reliability were related to an SLP's planning of intervention; items with the lowest evidence of reliability were related to an SLP's management of the session and rapport with students.

Despite some limitations, it was concluded the SLP performance evaluation instrument showed initial evidence of being able to evaluate SLPs fairly, accurately, and with perceived credibility from the district SLPs.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter One: Introduction

Recent school reform and educational policy have emphasized that teacher quality is a system of individual accountability for professional development and improvement (Every Student Succeeds Act, 2015). Measures of teacher effectiveness have been positively associated with student achievement (Kane & Staiger, 2012). Furthermore, systematic support to improve teaching practices has been shown to help educators meet the needs of students and articulate the process of change for improving student outcomes (Pianta & Hamre, 2009), even in the face of initial hesitation by staff to engage in a performance evaluation process (Corcoran, 2016). To date, however, educator evaluation research has primarily focused on general education standardized testing areas such as reading and math (e.g., Ho & Kane, 2013) with some focus on general education non-standardized testing areas such as science (Schultz & Pecheone, 2014). Very little emphasis has been placed on teachers and other professionals who provide services to unique student populations such as students with disabilities (Goe & Holdheide, 2011).

While states, districts, and practitioners recognize the need for valid and reliable performance evaluation systems for evaluated professionals within schools, challenges exist when evaluating non-classroom educators who have unique roles and responsibilities outside of a teacher (Holdheide, Hayes, & Goe, 2014). One of the primary challenges is the fact that instruction for non-classroom educators is often specific to individualized student needs. While some professional standards and best practices of non-classroom educators overlap with teachers, others have less overlap. In addition to individualized instruction, non-classroom educators typically have additional responsibilities that are essential to their job, such as completing individualized educational plans, conducting individualized evaluations, attending mandatory meetings, maintaining compliance with legal mandates, and modifying and

adapting curriculum for unique needs. To most effectively evaluate the performance of non-classroom educators, it may be necessary to modify or create a performance evaluation instrument that represents their unique roles, responsibilities, and best practices for their professions.

Like all students, students with communication disorders depend on high-quality services in order to achieve their goals. Educational speech-language pathologists (SLPs) prevent, assess, and provide intervention for communication disorders in schools (American Speech-Language-Hearing Association; ASHA, 2010). Performance evaluation systems that assess speech-language pathology professional competencies can provide important feedback to SLPs, which in turn, can assist clinicians in identifying areas of professional strength and growth, become more mindful of their practice, and engage in learning opportunities to close any gaps between current and desired performance (Hancock & Brundage, 2010). Additionally, performance evaluation provides important information to speech-language pathology administrators about the professional needs of their speech-language pathology staff. Access to such information may increase the likelihood of the delivery of high-quality services to students.

Given their unique roles and responsibilities in the schools, a performance evaluation instrument for SLPs must represent effective practices for serving children with communication disorders and all associated responsibilities, including the prevention of communication disorders, and compliance with district, state, and federal guidelines. Given the dynamic roles of educational SLPs, a performance evaluation instrument must be rigorous and systematic, but also flexible and sufficiently comprehensive to reflect roles and responsibilities across a wide range of students and educational contexts.

Rubric-based Performance Evaluation Instruments

There are several ways to evaluate performance and provide professional feedback to educators. For professionals who provide direct instruction to students, rubric-based observations of teaching are frequently used evaluation instruments (Center on Great Teachers and Leaders, 2014). In rubric-based observations of instruction, evaluators observe educators, collect evidence pertaining to a set of agreed-upon standards, and provide feedback to those evaluated based on a scaled continuum of performance for each standard (Pianta & Hamre, 2009). The structured professional conversations about the educator's performance, against the standards of performance, are intended to guide educators in enhancing their practice and in turn, improve student learning (Danielson, 2007).

In the field of speech-language pathology, the Performance Assessment of Contributions and Effectiveness (PACE, 2014) is an evaluation instrument specifically designed for school-based SLPs. The PACE was developed by the United States national certification agency for SLPs, the American Speech-Language-Hearing Association, and evaluates a SLP's performance in nine areas of practice based on the *Roles and Responsibilities of Speech-Language Pathologists in Schools* (ASHA, 2010). Each item on the PACE is not placed in a rubric of scaled performance levels (for example, performance that might indicate a skill level is *Unsatisfactory, Developing*, *Proficient*, or *Distinguished*); rather, the evaluator provides narrative comments on the SLP's performance based on the evidence collected. As such, while flexible and sufficiently comprehensive, it is currently unclear whether the PACE instrument produces reliable results.

If, however, a performance evaluation instrument for SLPs will be used to guide professional growth, then it will be necessary to have information about a performance evaluation instrument's reliability and validity. This is particularly true for professionals who wish to obtain a baseline of performance and use the performance evaluation

instrument to measure the effects of their individual professional growth in performance areas. A detailed rubric with specific performance levels for PACE items could certainly be designed and examined for evidence of reliability. But, school-based SLPs are not just members of their national certification agency (ASHA); they are important members of their local school districts and thus, it may be equally important a performance evaluation instrument for SLPs is consistent with a shared sense of local responsibility for school success (Holdheide et al., 2014).

With this in mind, several school districts have designed rubric-based performance evaluation instruments for their SLPs that are dually consistent with the PACE and local district initiatives (Center on Great Teachers and Leaders, 2014). To date, there are no published data on the reliability and validity of district-designed performance evaluation instruments for SLPs.

This Present Study: An Exploration of Evaluating SLPs

The present study examines evidence of reliability and validity for an experimental performance evaluation instrument specifically designed for school-based SLPs. The performance evaluation instrument (Appendix A) is a rubric-based evaluation of performance using an observation of instruction, and an examination of professional artifacts. This experimental performance evaluation instrument was developed to meet state regulations for teacher evaluation at a district level in 2014-2015.

This 2014-2015 performance evaluation instrument provided feedback to SLPs on 27 behaviors used in speech-language pathology services in Birth through age 21 school settings. The 27 behaviors were organized into four domains: Planning and Preparation of Service (six items), Climate of Service (four items), Implementation of Service (six items), and Professional Responsibilities, Due Process Documentation, and Case Management (eleven items). Domains 2 and 3 were evaluated based on an observation of service while Domains 1 and 4 were evaluated based on an examination

of professional artifacts (e.g., student work samples, professional documents, evidence of professional development). There were four scaled performance levels for each item on the instrument including *Requires Attention*, *Developing*, *Proficient*, *or Exemplary* levels, and each scale level was defined behaviorally using specific indicators of performance. The evaluation process occurred one time per SLP during the 2014-2015 school year.

The aim of the present study is to examine the preliminary validity and reliability evidence for this experimental performance evaluation instrument using archived data from 111 SLPs who participated in the evaluation process in 2014-2015. There are three research aims for this study:

1. To describe the performance of educational speech-language pathologists on an experimental performance evaluation instrument;
2. To determine the extent to which items on a speech-language pathologist performance evaluation instrument demonstrate construct and face validity; and
3. To determine the extent to which items on a speech-language pathologist performance evaluation instrument demonstrate agreement and produce reliable results.

This study is organized as follows: (a) Chapter One provides the rationale and overview of the study; (b) Chapter Two describes the literature used to ground the study in current theory and research; (c) Chapter Three consists of the study design; (d) Chapter Four describes the results of the study; and (e) Chapter Five discusses the interpretation and significance of the results, as well as future directions.

Chapter Two: Review of Literature

The overall purpose of this literature review is to review and highlight previous studies that inform the task of designing and validating an experimental performance evaluation tool for educational speech-language pathologists (SLPs).

This literature review covers the following content areas: (a) the intent and significance of performance evaluation for educational professionals; (b) research on the validity and reliability of performance evaluation for general education teachers; (c) research on the validity and reliability of performance evaluation for special education teachers; (d) position papers on performance evaluation for specialized support personnel in schools, such as SLPs; and (e) research on SLP quality and current descriptive tools for the performance evaluation of SLPs in schools.

Intent of Performance Evaluation for Educators

At its core, the intent of performance evaluation in schools is to assess and promote instructional quality of educators. The definition of "educator" varies from state to state, but commonly includes all education professionals whose roles and responsibilities include direct instruction to students including general education teachers, special education teachers, and other specialized support personnel, such as SLPs, occupational and physical therapists, social workers, school counselors, nurses, and other professionals (Center on Great Teachers and Leaders, 2014).

Results of performance evaluation in schools may be used summatively, for accountability purposes, or formatively, for professional development purposes. While policymakers have traditionally placed greater emphasis on the accountability function of performance evaluation, empirical research suggests the formative function of performance evaluation actually provides greater insight on the instructional activities and pathways that promote student learning (Papay, 2012; Reinhorn, Johnson, & Simon, 2016).

Individual states or districts design evaluation systems for their educators, typically using multiple measures from six common categories of performance evaluation measures (Center on Great Teachers and Leaders, 2014): (a) observations of instruction; (b) review of instructional and professional artifacts; (c) educator self-report of instructional and professional practices; (d) parent and student report; (e) value-added modeling (VAM), which provides an estimate of an educator's contribution to student test growth; and (f) other non-VAM student outcome measures, such as progress towards classroom or student learning objectives.

Two of these six measures, VAM and observations of instruction, have received relatively more attention by policymakers, researchers, and educators in the design and implementation of performance evaluation systems (Hallinger, Heck, & Murphy, 2014). In VAM, student test performance is predicted by past test performance, student background characteristics, and/or classroom or school characteristics. The contribution, or value added, of an individual educator is estimated by comparing his/her students' actual to predicted performance. As such, VAM has a specific focus on test-score growth. What may be conceptually easy to understand in VAM -- the concept of isolating an educator's contribution to test-score growth -- is often methodically challenging in practice because students are not typically randomized into classrooms (i.e., some teachers have certain types of students), students often learn from multiple educators, technical properties of tests themselves impact growth estimates, and only teachers in tested areas receive VAM scores (Papay, 2012).

In observations of instruction, states or districts adopt or develop a set of standards and a rubric that describes a continuum of performance for each standard. Trained evaluators observe educators, collect evidence to the standards and rubric, and provide performance feedback based on the evidence (Pianta & Hamre, 2009). For many observation instruments, standards are clustered into domains, such as practices

related to the planning and preparation of instruction, practices related to the delivery of instruction, and practices related to school-wide professional culture. Although domains are designed to be distinct, they are often linked by a theoretical framework and empirically related to one another. For example, the Framework for Teaching (or, FFT; Danielson, 2007) is a performance evaluation instrument that measures four domains of teaching, divided into 22 elements, based on the theoretical framework of constructivism. Although the practices in the four FFT domains are measured separately, data from Lash, Tran, and Huang (2016) showed that they are correlated: a teacher's planning and preparation of instruction (Domain 1) was related to his/her teacher's classroom management (Domain 2), delivery of instruction (Domain 3), and completion of professional responsibilities (Domain 4).

Performance observation instruments may be specific to grade and/or content areas such as English Language Arts (e.g., Grossman et al., 2010), mathematics (e.g., Hill et al., 2008), or science (e.g., Schultz & Pecheone, 2014). More commonly, though, school districts employ the use of generic observation instruments that span across grade levels and content areas (Kane & Staiger, 2012) such as the FFT (Danielson, 2007), Marzano's Teacher Evaluation Model (Marzano & Toth, 2013), or the Classroom Assessment Scoring System (CLASS; Teachstone, 2013). Using a rubric that spans across content and grade levels is believed to promote a shared understanding of teacher effectiveness and a common language for professional improvement.

Similar to VAM, the measurement and psychometric issues of observational instruments can be of great consequence (Pianta & Hamre, 2009). In any observation instrument, the definition of educator quality can be difficult to operationalize (Fenstermacher & Richardson, 2005). Moreover, sources of bias can be threats to validity (Papay, 2012; Park, Chen, & Holtzman, 2014). The reliability of scores may be affected by the sampling of lessons, differences among raters, data collection mode (live

or videotaped), and the level of inferencing needed to score individual instrument items (Hill, Charalambous, & Kraft, 2012; Ho & Kane, 2013; Praetorius, Pauli, Reusser, & Rakoczy, 2014). For example, Hill et al. showed the reliability of scores was lower for some (but not all) items on an observation tool for mathematics teaching when evaluators scored the first 30 minutes of a lesson compared to the entire lesson and reliability was higher for most (but not all) items when three mathematics lessons from a teacher were observed compared to two lessons.

The intent of any performance evaluation system--whether it includes only VAM, only observations, or a combination of measures--is to offer educators a systematic way to assess and develop their practice using a common framework and clear statements of defined performance. Theses structured conversations about performance are intended to guide educators in enhancing their practice and in turn, increase student outcomes (Danielson, 2007).

Significance of Performance Evaluation for Educators

Performance evaluation systems for educators have been viewed as an important strategy for school improvement, in part, because of their ability to quantify the contribution of individual teachers on student outcomes. Performance evaluation research has shown classroom teachers account for a significant amount of variance in student achievement and are the most important school-level factor in predicting student outcomes (Ferguson, & Danielson, 2014; Kane, McCaffrey, Miller, & Staiger, 2013; Kane & Staiger, 2008; Rivkin, Hanushek, & Kain, 2005; Staiger & Kane, 2014).

Additionally, performance evaluation has received attention for its ability to improve instructional practices and as a result, increase student outcomes. In the past decade, competitive federal funding that sought to improve teacher evaluation systems (e.g., Race to the Top; RTTP, US Department of Education, 2009) led to an unprecedented amount of information about the impact of performance evaluation

systems on student outcomes (Center on Great Teachers and Leaders, 2014). One of earliest studies in this new wave of RTTP research was by Taylor and Tyler (2012). They studied the impact of a district's performance evaluation system on mathematics teaching for 105 midcareer teachers. The performance evaluation system consisted of four observations of instructions per year. Three of these observations were conducted by a peer observer and one observation was conducted by an administrator, but all used the FFT (Danielson, 2007) as the observation rubric in which to organize the feedback and guide the professional conversations. Taylor and Tyler showed that students who were taught by a teacher who had participated in the district's evaluation system made an average gain of 4.5 percentile points in math achievement scores, compared to a similar student taught by same teacher before the district's evaluation system was implemented. Moreover, the changes in the teacher's mathematics teaching, based on the performance evaluation feedback, extended to the year after the evaluation.

Similarly, in their study of 35 elementary and middle charter schools in New York City, Dobbie and Fryer (2013) showed that schools who provided teachers with specific instructional feedback based on observations of instruction had higher math and English Language Arts standardized test score gains than schools who did not provide teachers with specific instructional feedback. These gains were statistically significant, even after controlling for individual student factors (e.g., baseline test scores, free lunch status, gender, race, grade, lottery admission into charter school).

In another study, Allen, Hafen, Gregory, Mikami, and Pianta (2015) showed that middle- and high-school teachers who participated in a coaching intervention program, based on specific performance evaluation ratings, produced significantly higher student achievement scores compared to teachers in a control condition. In their study, 87 teachers in 5 different middle and high schools were randomly assigned to a treatment or control condition for the duration of a two-year study. The treatment condition

included web-based and in-person coaching and the control condition included business-as-usual, in-service training. Student outcomes were defined as standardized test score gains at the end of year two, controlled for prior levels of achievement. Pre-treatment differences between treatment and control conditions were non-significant for all student and teacher characteristics, including baseline student test scores. At the end of the two years of targeted coaching, results showed test score gains were significantly higher for teachers in the treatment condition compared to the control condition.

Results from Allen et al. (2015) were consistent with a later study by Papay, Taylor, Tyler, and Laski (2016). Papay et al. showed that when low-performing teachers were paired with high-performing teachers, based on ratings from performance evaluation, student achievement scores in both reading and math improved. In their study, 90 low-performing teachers in 16 different schools were randomly assigned to a treatment or control condition. The treatment condition included peer coaching with a high-performing peer to improve teaching skills identified from performance evaluation ratings. The control condition consisted of the district's business-as-usual professional development provided to the low-performing teachers. Pre-treatment differences between treatment and control conditions were non-significant for baseline student achievement scores. At the end of the academic year, student test scores were significantly higher for teachers in the treatment condition compared to the control condition. The student gains were highest when low- and high-performing teachers were matched on specific skills identified from the performance evaluation versus being matched on general, non-specific performance such as matching only based on a teacher's grade level (e.g., 4th-grade teachers are only matched with other 4th-grade teachers) or overall performance (e.g., overall low-performing teachers are only matched with overall exemplary teachers).

In the area of special education teaching, Johnson (2015) provided preliminary evidence regarding the impact of performance evaluation on instruction for students with disabilities. Johnson used a performance evaluation tool to plan targeted coaching for teacher candidates in a special education master's program. The performance evaluation tool, Recognizing Effective Special Education Teachers (RESET), was specifically designed to provide actionable feedback on the implementation of evidence-based practices for students with disabilities. Using observations of their instruction and the RESET, the teacher candidates received feedback from supervising university faculty on the implementation of evidence-based instructional practices. Given the modest sample size of ten teacher candidates and two supervising faculty, outcomes in the study were descriptive in nature, and were based on survey feedback. Survey feedback showed candidates responded positively to the pilot program, indicating the explicit criteria of performance on the RESET helped them translate research to practice. Faculty responses were also positive, indicating the RESET provided an analytic and reflective focus to the observation process and contributed to revisions of graduate coursework to include greater emphasis on key evidence-based practices.

If performance evaluation is to improve instructional practices of educators and in turn, student outcomes, performance evaluation systems must not only document educator quality but also develop and support them. For professionals who provide instruction to students, performance evaluation provides feedback on the basic definition of successful teaching (Fenstermacher & Richardson, 2005, pp. 3-4) which is defined:

(1) There is a person, T, who possesses some

(2) content, C, and who

(3) intends to convey or impart C to

(4) a person, S, who initially lacks C, such that

(5) T and S engage in a relationship for the purpose of S's acquiring C and

(6) S acquires C to some acceptable or appropriate level

Performance evaluation feedback enables educators to leverage their areas of strength and improve their areas of weaknesses in order to increase learning for their students (Papay, 2012; Pianta & Hamre, 2009).

When performance evaluation does not provide actionable feedback to practitioners that is specific and meaningful, the potential for professional growth is lessened, as is the potential to improve student outcomes. In a review of performance evaluation research from 1997-2013, Hallinger, Heck, and Murphy (2014) showed that student outcome gains were non-existent or minimal when school administrators emphasized the summative, "compliance" function of performance evaluation over the formative function. Furthermore, even when school administrators approached performance evaluation from a formative framework, they were not always equipped to provide tailored feedback and as result, teachers were less likely to make instructional changes (Kraft & Gilmour, 2016; Rigby et al., 2017; Vickers, 2015) that benefitted all students (Steinberg & Sartain, 2015).

In order for a performance evaluation tool to assess and develop educator quality, it should be valid (i.e., it measures what it purports to measures), reliable (i.e., it gives you the same answer regardless of the evaluator), free from bias (it gives you the same answer if you use it with diverse groups of educators) and have capacity to drive individual and organizational change (Papay, 2012). To date, the development and validation of performance evaluation systems in schools has primarily focused on general education teachers, with some emphasis on special education teachers. Identifying and assessing the major findings from these two bodies of research can inform the development and validation of performance evaluation tools for other non-classroom professionals, including the focus of this study, school-based SLPs.

Performance Evaluation for General Education Teachers

The history of performance evaluation research in general education dates back to the 1940's, although the 1960's were the primary period in which researchers began to study the link between teaching quality and student learning (see Blanton et al., 2006, for a brief review). This work initially focused on process-product observational instruments, which sought to identify specific teacher actions responsible for producing student learning. This product-process approach extended into the 1970's and beyond, but additional variables affecting student learning were also identified, including the complex relationships between teachers, classrooms, and schools. Teacher observation tools were then enriched to include the classroom and school contextual factors that were proposed to help students learn (e.g., Englert, Tarrant, & Mariage, 1992). Researchers studying teaching quality made a distinction between *good teaching*, which encompasses the teacher characteristics to meet the basic expectations of the profession (e.g., holding degrees, receiving content training, upholding the standards of the field, applying appropriate pedagogy, and other teacher attributes and beliefs about learning) and *effective teaching*, which encompasses the instructional methods and activities that result in student learning or achievement (Berliner, 1987). The 2000's were marked by an increased emphasis on standardized teacher observation instruments, given the emphasis of performance evaluation in federal grants (e.g., Race to the Top; RTTP, US Department of Education, 2009), and several research projects were initiated to help schools better assess and develop instructional quality using performance evaluation tools.

The Measures of Effective Teaching Project (MET; Gates Foundation, 2013) may be the largest United States research project with the stated purpose of helping schools implement performance evaluation. With over $45 million in funding, two dozen academic and organizational partners, and data from 3,000 teachers and 100,000 students, the MET project has produced large databases of teaching practices that can

be accessed for empirical analyses (Archer, Kerr, & Pianta, 2014). Internationally, there are similar large-scale research projects organized for the same purposes (e.g., Thomas, 2001; van de Grift, 2014). In general, these large-scale research projects have focused on tested areas such as reading and math.

Several major findings about performance evaluation systems and instruments have emerged from these large-scale research projects. These findings are summarized below:

1. Performance evaluation systems that include a combination of performance evaluation measures provide a more valid and reliable estimate of effective instruction than any single performance evaluation measure (Kane et al., 2013; Kane & Staiger, 2012; Walkington & Marder, 2014).
2. The validation of performance evaluation instruments should involve a broad set of psychometric examinations, and should not be limited to correlating teacher performance on these instruments to student test score gains, as in done in VAM (Hill et al., 2012a; Pianta & Hamre, 2009).
3. Reliability of performance evaluation instruments may be impacted by multiple sources of variance (e.g., differences in raters, lessons, instruments themselves, and/or sampling of data). Moreover, the interpretation of reliability examinations depends on the analytic strategies employed to attend to these potential sources of variance (Hill et al., 2012a; Hill, Charalambous, & Kraft, 2012; Ho & Kane, 2013; Joe, McClellan, & Holtzman, 2014; Praetorius, Pauli, Reusser, & Rakoczy, 2014).
4. Bias in performance evaluation instruments is not limited to instruments that rely on human judgments for scoring (e.g., observations of instruction), but can occur in other instruments where the assignment of students to teachers is not random

(i.e., certain types of students tend to be grouped with certain teachers; Hallinger et al., 2014; Papay, 2012; Steinberg & Sartain, 2015; Kane et al., 2013).

5. Performance evaluation instruments, to date, that appear to have the most promising evidence of adequate validity and acceptable reliability for assessing and developing teaching skills include standardized observations of instruction and student perception surveys (Ferguson, & Danielson, 2014; Pianta & Kerr, 2014; Raudenbush & Jean, 2014; Ruzek, Hafen, Hamre, & Pianta, 2014), although both are typically combined with student achievement measures in a performance evaluation system for general education teachers (Center on Great Teachers and Leaders, 2014).

School districts and the public have also been alerted to remaining needs in performance evaluation research from these large-scale empirical projects. To begin, the bulk of performance evaluation research has been limited to teachers in tested grades (typically 4th-8th reading and math) but nationally, teachers in tested areas represent only 31% of the educator population (Prince et al., 2009). Secondly, performance evaluation research has primarily been limited to students without disabilities (Sledge & Pazey, 2013), but 12.9% of all students in America's schools receive special education services (Snyder & Dillow, 2015).

Given 12.9% is a sizable percentage of students, the Council of Exceptional Children (2012), Center for Great Teachers and Leaders (Holdheide et al., 2010), and researchers in special education teaching (e.g., Jones & Brownell, 2014) have provided guidelines to states and districts for developing performance evaluation systems for special educators. Furthermore, at least one state (Idaho) is in the validation stages for a performance evaluation tool specifically designed for special education teachers (Semmelroth & Johnson, 2014).

The next section discusses performance evaluation systems in special education teaching, with a specific focus on observation tools as performance evaluation measures. Following this section, I discuss the current status of performance evaluation for other non-classroom, "caseload" staff who also serve students with disabilities, such as SLPs.

Performance Evaluation for Special Education Teachers

The goal of any performance evaluation system for professional development purposes is to help teachers understand the types of instructional actions that predict student learning and provide professional feedback to enhance their practice and in turn, promote student learning (Danielson, 2007). Developing and validating performance evaluation instruments to achieve this goal in general education has proven to be quite challenging (Kane, Kerr, & Pianta, 2014). This challenge may be even greater when the teachers of focus include those who work under a variety of conditions to meet the specialized, individual needs of a heterogeneous group of students with disabilities (Johnson & Semmelroth, 2014; Jones & Brownell, 2014).

Performance evaluation measures for unique student populations have primarily focused on standardized observations of instruction, in part because the standardized achievement tests used in general education may be neither available nor appropriate for special education use. Johnson and Semmelroth (2012; 2014) and Sledge and Pazey (2013) discussed the psychometric and methodological challenges of using standardized achievement tests as an indication of special education teaching effectiveness, noting the number of special education students associated with a special educator is often too few to be used in quantitative analyses of student achievement outcomes; the growth rate of academic skills for students with disabilities may be variable and lead to inaccurate calculations of value-added scores; students with disabilities may have accommodations and modifications for testing on their Individual

Education Plans, potentially invalidating the standardized testing conditions; and in general, the use of standardized tests to measure growth for students with moderate-severe cognitive disabilities is questionable at best. Standardized observations of instruction have been a promising tool for performance evaluation in special education because they have the potential to capture the unique roles, responsibilities, and expectations of expertise for special education teachers, despite significant variations in students and contexts served (Jones & Brownell, 2014).

Too often, however, special educators are evaluated on an observation instrument that is not specific to the instructional actions known to be effective for students with disabilities. In a survey of school district special education administrators across 51 states (Holdheide et al., 2010), 84.1% of the 1,100 respondents reported special education teachers are required to have knowledge, skills, and expertise that general education teachers do not have, yet only 26% of the respondents' school districts allowed for a modified or separate performance observation measure for special education teachers. Additionally, when standardized observation measures were used, only 12.4% of the respondents' school districts provided explicit training for evaluators of special education teachers.

Two lines of research have focused on the development and validation of standardized observation instruments for the evaluation of special education teaching. One line of work includes the proposal to modify current general education observation tools for special education use (e.g., Jones & Brownell, 2014); a second line of work includes the development of unique observation tools specifically designed for special education teachers (Semmelroth & Johnson 2014).

Modifying current general education performance evaluation tools for special education use has been attractive for several reasons: (a) having a shared understanding of effective teaching across general and special education promotes a

shared vision of improved student outcomes; (b) training multiple raters on a single instrument is more cost-effective than training multiple raters on multiple instruments; and (c) modifying an observation tool that has been substantiated for use in general education research may facilitate a quicker validation process for use in special education (Holdheide et al., 2010; Jones & Brownell, 2014).

Jones & Brownell (2014) examined a generic teacher standardized observation evaluation tool (FFT; Danielson, 2007) against a definition of effective special education teaching. Based on an aggregate of evidence-based practices from an extensive literature review, Jones and Brownell defined effective special education teaching as being (a) explicit; (b) intensive; (c) cohesive; (d) engaging; (e) responsive; and (f) focused on essential concepts, strategies, and skills. When the FFT was compared against this definition, some of the effective special education practices were represented on the FFT while other practices were not represented. The most notable difference was explicitness of instruction; whereas the FFT has some emphasis on explicit instruction, evidence-based practices for special education have a high emphasis on explicit, clear, systematic instruction from a teacher. An additional area of divergence included the reliance of student actions as evidence of teaching effectiveness. On the FFT, *distinguished* (highest level) performance of a general education teacher may be evidenced by students' unprompted verbal initiations and responses. For some teachers of students with disabilities, the quality of instruction may be inferred from student responses, but for other teachers of students with disabilities, *distinguished* instruction may not be so easily inferred from student responses.

If general education performance evaluation tools are to be modified for special education purposes, Holdheide et al. (2010) provides these recommendations:

1. Include special education administrators and teachers when modifying or designing performance evaluation instruments.

2. Differentiate the evaluation process for special education teachers where appropriate. In particular, make sure the pre-observation conference provides opportunities for special educators to provide information about the unique, specialized instruction they provide and the outcomes used to measure student progress.
3. When using a common performance evaluation rubric for all teachers, provide explicit examples of criteria/expectations for special education teachers based on their evidence-based practices and roles/responsibilities as special educators.
4. Establish evaluator training that includes effective practices for special education teaching and/or consider using peer-to-peer evaluators that are matched to specific special education disciplines.
5. Ensure the results of the modified framework have instructional value to special educators and provide a path to identify individual, professional development needs.
6. Continue to evaluate the validity and reliability of performance evaluation tools that are modified for special education purposes.

As an alternative to modifying general education observation tools, researchers in at least one state are in the stages of validating a performance evaluation tool uniquely designed for special educators. The Recognizing Effective Special Education Teachers (RESET) is a performance evaluation tool developed in Idaho to measure the effectiveness of special educators (Johnson & Semmelroth, 2012). The RESET was developed based a definition that an "effective special education teacher is someone who is able to identify a student's needs, implement evidence-based instructional practices and interventions, and demonstrate student growth" (Johnson & Semmelroth, p. 76).

The RESET observation tool consists of three parts, including an evaluation of the teacher's: (a) lesson overview; (b) implementation of specific lesson components, which includes the features of specific evidence-based practices; and (c) lesson summary. Each item within the three parts is evaluated on a four-point scale of *Unsatisfactory*, *Basic*, *Proficient*, and *Distinguished*. Depending on the specific evidence-based practices implemented, there are approximately 30 - 70 items on the RESET. To date, the RESET is not yet in widespread use outside of experimental studies (Lawson, 2015; Semmelroth & Johnson, 2014) and to the author's knowledge, there are no other modified or unique performance evaluation tools for special educators.

Performance Evaluation for Specialized Support Personnel

Special educators have unique roles and responsibilities within schools, not unlike the specialized roles and responsibilities of non-classroom educators. The term, "Specialized Instructional Support Personnel" (SISP) describes the group of education professionals who have unique roles and responsibilities outside of the classroom. In the Every Student Succeeds Act (2015), the term SISP is defined as: "(i) school counselors, school social workers, and school psychologists; and (ii) other qualified professional personnel, such as school nurses, SLPs, and school librarians, involved in providing assessment, diagnosis, counseling, educational, therapeutic, and other necessary services (including related services as that term is defined in section 602 of the Individuals with Disabilities Education Act (20 U.S.C. 1401) as part of a comprehensive program to meet student needs.''

The Center for Great Teachers and Leaders provides two position papers on general considerations for the performance evaluation of SISP (Holdheide et al., 2010; Holdheide, Hayes, & Goe, 2014), as many states must include SISP in their performance evaluation systems to meet statutory or regulatory requirements (Center for

Great Teachers and Leaders, 2014b). In general, these two positional papers state the considerations for SISP performance evaluation are similar to the considerations for special education teaching performance evaluation. Key considerations include using differentiated performance evaluation systems for SISP as a way to ensure that performance evaluation is effective in promoting professional change for SISP. Both teachers and SISP want to provide high-quality instruction, but they may vary in the skills and evidence-based practices needed for their respective professions and/or ways these skills are demonstrated for the students they serve. Secondly, aligning performance evaluation to SISP national association standards ensures a more fair assessment of professional skills. At the same time, it may be equally important for SISP performance evaluation to promote a shared sense of responsibility for student learning at a local district level. Finally, to increase SISP confidence an evaluation system fairly and accurately captures SISP performance, trained SISP peer evaluators may be an additional or alternative option to principals or district administrators. When using this approach, peers are trained to evaluate their SISP colleagues using a discipline-specific rubric, which enhances evaluator credibility, but also leverages the expertise of SISP leaders and encourages a culture of learning for both evaluators and SISP being evaluated.

Performance Evaluation for SLPs

Educational SLPs are a subgroup of SISP who prevent, assess, plan, and provide instruction/intervention for communication disorders in educational contexts (American Speech-Language-Hearing Association; ASHA, 2010). A communication disorder is defined as a developmental or acquired impairment in one or more areas of speech, language, fluency, swallowing, and voice/resonance (ASHA, 1993). SLPs serve approximately 8% of the student population nationally (Black, Vahratian, & Hoffman, 2015). There is significant variation in the contexts in which SLPs serve students. In

some states, educational SLPs serve students at ages Birth-22 while in other states, SLPs serve students in K-12 settings only. Students may receive speech-language services in their homes, community settings, general education classrooms, special education classrooms, and/or speech-language rooms. Services may include direct instruction, supports, and/or accommodations provided by the SLP only, in collaboration with families or other team members, or through consultation with families or other team members. Students with a communication disorder may demonstrate one or more speech-language disabilities as the only area of disability, or their disabilities may be secondary to other conditions such as Autism, cognitive delay, or other health impairments. The strongly interdisciplinary nature of SLPs requires them to serve schools in multiple capacities across a wide range of educational contexts.

The individualized instruction across a wide range of educational contexts makes evaluating school-based SLPs challenging. An evaluation instrument for school-based SLPs needs to be comprehensive, yet flexible to accommodate the wide heterogeneity in services. Additionally, an evaluation tool for school-based SLPs must represent the duties and responsibilities beyond direct services provided to students, such as compliance with legal mandates, engagement with families, and collaboration with other professionals. Finally, an evaluation tool for school-based SLPs should be consistent with national standards for the profession yet contribute to a shared sense of local school success (Holdheide et al., 2014).

The Every Student Succeeds Act (2015) and Individuals with Disabilities Education Act (2004) recognize the importance of SLPs in schools. Services provided to improve communication skills of all students is at the core of the speech-language pathology and one way to ensure students receive the highest quality of services is through a valid performance evaluation system.

To date, there are no published research studies on performance evaluation for practicing SLPs. There is, however, a small but growing research base that examines the role of the practicing clinician on student outcomes. While not performance evaluation research per se, findings from these studies can shed light on the features of SLP quality that may be important to include on a performance evaluation instrument for school-based SLPs. The first study to mention was conducted by Kamhi (1995). Kamhi used a series of descriptive survey studies that surveyed practicing clinicians' beliefs about factors they felt were important for providing effective therapy. Based on survey responses, Kamhi proposed a model of clinical expertise that included four components: (a) knowledge base in speech-language development, disorders, and clinical methods; (b) procedural and problem-solving skills for speech-language clinical methods; (c) interpersonal skills and attitudes, including rapport with clients, confidence, adaptability, enthusiasm, and interest; and (d) on-going self-monitoring skills to assess competency in and engage in lifelong learning for the other three areas. When asked to identify which skills were most important in effective therapy, clinicians generally indicated knowledge and procedural skills were just as important as interpersonal skills and attitudes.

More recently, Ebert and Kohnert (2010) conducted a similar descriptive survey study to examine which features of speech-language therapy were regarded as important by second-year SLP graduate students and practicing SLPs in influencing treatment outcomes. In their two-phase study, Ebert and Kohnert first analyzed themes from an open-ended survey about what the participants judged to be effective treatment for clients. Information from this first study was then used to create a survey for a second study of practicing SLPs, only, who were asked to rate the importance of 25 clinician qualities on a scale of 1-5 (1=Negligible impact on therapy outcome; 5=Very large impact on therapy outcome). Approximately 72% of the participants in study two

were school-based SLPs and nearly all provided a wide range services to students. Results showed clinicians ranked three qualities to be most important: the clinician's rapport with the client; communication between the clinician and the client; and how well the clinician places therapy in a functional context. Other important qualities, with slightly lower median ratings, included the clinician's willingness to change intervention goals and activities; clinician's theoretical framework for understanding the disorder; extent of communication between clinician and client's family; degree to which clinicians uses principles of evidence-based practice; and how often the clinician reconsiders the client's progress. The quality with the lowest median rating included the amount of time clinician spends filling out supporting paperwork.

In 2017, Ebert continued her work on examining the potential impact of SLP qualities on treatment outcomes by exploring the influence of the clinician-client relationships on intervention progress for school-age language impairment. In her study, Ebert hypothesized a positive collaborative relationship between an SLP and his/her client and their caregiver would predict treatment outcomes for the client. Ebert measured the clinician-client relationship using a 12-item rating scale adapted from counseling psychology. SLPs, children, and the children's caregivers completed the adapted rating scales immediately prior to and following a four-month period of language impairment intervention. Treatment outcomes included caregiver and SLP perceptions change on an informal Likert scale of overall treatment progress from *Far less progress than expected* to *More progress than I expected* as well as change on a standardized checklist of communication skills and the percent of treatment sessions attended. In this pilot study, results showed that a caregiver's positive view of the clinician-client relationship was associated with positive change on all three outcome measures. Additionally, a SLP's positive view of the clinician-client relationship was associated with change on the informal measure of overall progress.

In a slightly different type of study, Justice and colleagues examined factors of SLP quality as part of a larger, three-year multi-cohort study called Speech-Therapy Experiences in Public Schools (STEPS; see Farquharson, Tambyraja, Logan, Justice, & Schmitt, 2015, for a description of STEPS). This study was designed to examine key active ingredients of speech-language pathology intervention for school-age children with language impairment and resulted in a large database of information from 73 educational SLPs and nearly 300 students. In addition to demographic and work-related variables, measures of SLP quality were planned as potential predictor variables in an analysis of student outcome growth. Justice and colleagues defined SLP quality as scores on the Classroom Assessment Scoring System (CLASS; Teachstone, 2013), a commonly-used observation tool for assessing classroom teaching quality. To obtain CLASS scores, SLPs submitted videotaped intervention sessions. Given the CLASS observation tool was developed for teachers, it was slightly modified for application to SLPs (see Biancone, Farquharson, Justice, Schmitt, & Logan, 2014, for description of modifications). Trained coders scored videotaped SLP intervention session for three factors of quality using the CLASS observation tool: (a) *Emotional Support*, which includes dimensions of positive climate, sensitivity, and regard for student perspectives; (b) *Proactive Management*, which includes dimensions of behavior management, productivity, and instructional learning formats; and (c) *Instructional Support*, which includes the extent of concept development, quality of feedback to students, and language modeling provided by SLPs.

To date, the STEPS research team has published three studies that describe CLASS findings from their multi-cohort study (Biancone, Farquharson, Justice, Schmitt, & Logan, 2014; Farquharson et al., 2015; Schmitt, Justice, & O'Connell, 2014). Data analyses by Farquharson et al. demonstrated that approximately 8-%-12% of the variance in student outcomes could be explained by individual SLPs, which was similar

to the percentage of variance in student achievement explained by individual classroom teachers (Kane & Staiger, 2008), but their analyses were not conclusive about which CLASS factors, if any, explained differential outcomes for students. Schmitt et al. and Biancone et al. showed the CLASS factor of *Emotional Support* was a strong feature of the SLP intervention sessions, but it did not necessarily explain differential student outcomes, possibly because of the restricted range of high performance for SLPs. The CLASS factor *Proactive Management* was similarly strong for SLPs, but also showed a restricted range of high performance for SLPs, while the *Instructional Support* factor showed a larger range of performance for SLPs that may be possibly linked to student outcomes in future studies (Biancone et al.).

Taken together, the studies on SLP quality indicate that practicing SLPs believe technical skills, theoretical foundations of therapy, relationships with clients, flexibility and adaptability by the SLP, and functional therapy contexts are important to student outcomes. These factors may be important for districts to consider when developing performance evaluation instruments for SLPs. Additionally, empirical evidence from Ebert (2017) and from the STEPS research team suggests that individual SLPs do matter in intervention outcomes, just as individual teachers matter in student outcomes (e.g., Kane & Staiger, 2008), and thus, it is important to assess and develop SLP quality.

Special emphasis on EBP. Evidence-based practice (EBP) is defined as “the integration of best research evidence with clinical expertise and patient values" (Sackett, Straus, Richardson, Rosenberg, & Haynes, 2000). EBP is a core feature of the three primary responsibilities of a school-based SLP, which is to prevent, assess, and provide intervention for communication disorders in schools (ASHA, 2010):

- Prevention — SLPs are integrally involved in the efforts of schools to prevent academic failure in whatever form those initiatives may take; for example, in the implementation of EBP for Response to Intervention.

- Assessment — SLPs conduct assessments in collaboration with others that help to identify students with communication disorders as well as to inform intervention, consistent with current EBP for assessment and evaluation.
- Intervention — SLPs provide intervention that is appropriate to the age and learning needs of each individual student, to be selected through an evidence-based decision-making process.

The expectation of SLPs to implement EBP is based on accumulating evidence that the use of EBP is associated with positive student gains for students with communication disorders (see, for example, the compendium EBP systematic reviews on ASHA, 2016). Viewpoints of practicing clinicians certainly suggest EBP is an important factor in promoting positive therapy outcomes (Ebert & Kohnert, 2010) and a growing number of resources exist to support the translation of the EBP process into clinical practice (e.g., Ukrainetz, 2014; Williams, McLeod, & McCauley, 2010). The prominent role of EBP in the evaluation of special education teachers (Johnson & Semmelroth, 2012; Jones & Brownell, 2014) underscores the importance of EBP in service to students with disabilities. Given of the importance of EBP in promoting student outcomes, it appears particularly important to evaluate the extent to which an SLP engages in the EBP process in a performance evaluation instrument for school-based SLPs.

The next section briefly discusses the current performance evaluation tool for SLPs published by the national certifying agency for SLPs, the American-Speech-Language-Hearing Association.

The Performance Assessment of Contributions and Effectiveness (PACE). For school-based SLPs, standards of practice are codified in the *Roles and Responsibilities of Speech-Language Pathologists in Schools* (ASHA, 2010). In 2012,

these roles and responsibilities were used to develop (2012) and revise (2014) the Performance Assessment of Contributions and Effectiveness. The PACE is a performance evaluation tool intended to assist states/districts in the implementation of a performance evaluation system for school-based SLPs, as an alternative option to evaluating SLPs on a general education teacher rubric. Nine performance areas are evaluated on the PACE, including an SLP's ability to:

1. Demonstrate knowledge and skills in speech-language pathology and related subject areas (e.g., literacy) and implement services in an ethical manner.
2. Provide culturally and educationally appropriate services that are effective, engage students, and reflect evidence-based practice.
3. In partnership with the team, determine eligibility and recommend services that are compliant with state and federal regulations for children with IEPs.
4. Demonstrate ability to conduct appropriate comprehensive evaluations for students who may be experiencing a variety of communication disorders.
5. Use appropriate and dynamic service delivery methods consistent with the wide variety of individual student needs.
6. Collaborate with classroom teachers and other professionals to serve the needs of students in both general and special education.
7. Collaborate with families and provide opportunities for them to be involved in their student's speech-language pathology services.
8. Earn continuing education or professional development units sufficient to meet ASHA requirements for certification maintenance as well as state certification and licensing requirements.
9. Contribute to various building and/or district initiatives.

A variety of evidence types may be used to evaluate an individual SLPs' performance in the nine items above, including evidence collected from a portfolio of professional artifacts or direct observation of services to students.

The PACE provides an excellent comprehensive summary of educational SLP roles and responsibilities, but each item on the PACE is not scaled on a continuum of performance. That is, the nine PACE items do not contain levels of performance that would suggest an SLP's skills in a particular area require attention or are developing, proficient, or distinguished. As such, it is up to individual states/districts to determine performance levels and conduct reliability studies to determine if the PACE instrument produces reliable results for their SLPs.

In addition to the PACE, several states have allowed districts to modify teacher performance evaluation rubrics for the performance evaluation of their SLPs including the states of Colorado, District of Columbia, Florida, Rhode Island, Tennessee, and Alabama (Holdheide et al., 2014). There are no published data on the validity and reliability of these district-designed instruments, but all represent examples of SLP performance evaluation instruments that are consistent with the PACE as well as local district initiatives for school success.

Summary

Performance evaluation systems for teachers are specific to teaching; the instruments within these systems do not necessarily represent the work by other educational professionals within schools (Danielson, 2007). Like teachers, every educational profession establishes a common language to identify the roles, responsibilities, and expectations of expertise for practice. The purpose of a performance evaluation tool for educators is to assess and promote practitioner expertise and ensure instructional quality.

In order for a performance evaluation tool to ensure instructional quality, it should be valid (measures what it claims to measure), reliable (consistently provides the same answers), and free from bias (provides the same answers across diverse groups of people) (Papay, 2012). Without evidence of validity and reliability, a performance evaluation tool is limited in its ability to document professional strengths/needs, measure professional growth, and provide actionable feedback to practitioners (Pianta & Hamre, 2009).

To date, empirical evidence for the validity and reliability of performance evaluation systems for educators has largely focused on general education tested areas such as literacy and math (e.g., Measures of Effective Teaching; Gates Foundation, 2013) and the recommendations for performance evaluation systems for non-tested areas, such as speech-language pathology, have largely gone untested (Holdheide et al., 2010).

The present study aims to address the empirical gap in the performance evaluation literature by examining evidence for validity and reliability of an experimental performance evaluation tool developed for school-based SLPs. The tool was designed to be consistent with the PACE (ASHA, 2014) as well as the local district framework for teacher evaluation. If the performance evaluation instrument shows acceptable validity and reliability, it may serve as a systematic way for school-based SLPs to assess and develop their professional skills and in turn, promote student growth.

The present study had three primary research aims:

1. To describe the performance of educational SLPs on an experimental performance evaluation instrument;
2. To determine the extent to which items on an experimental SLP performance evaluation instrument demonstrated preliminary construct and face validity; and

3. To determine the extent to which items on an experimental SLP performance evaluation instrument demonstrated agreement and produced reliable results.

In the next section, I describe the methods and analytic strategies to address these research aims.

Chapter Three: Methods

The purpose of this study was to examine validity and reliability evidence for an experimental performance evaluation instrument for educational speech-language pathologists (SLPs). Data used in this study were de-identified, district program evaluation data, available in July 2015, archived from the 2014-2015 school year. This chapter describes the following: (a) the SLPs and evaluators included in the archived data set, (b) the experimental performance evaluation instrument used in the SLP evaluations, (c) procedures utilized by SLPs and evaluators to complete the evaluation process as well as the procedures to gather face validity and interrater reliability data, and (d) the descriptive and inferential statistical statistics employed to address the research aims.

Participants

SLPs. Performance evaluation data for this study were from 111 SLPs in a single, mid-size urban district. At the time of their evaluation, the SLPs provided demographic information to describe themselves.

The SLPs were primarily female (*n* = 108; 97%), Caucasian (*n* = 109; 98%), and held a Certificate of Clinical Competence from the American Speech-Language-Hearing Association (*n* = 107; 96%). All SLPs held a state license in educational speech-language pathology with a range of professional experience from 0 to 46 years (*M* = 17 years; *SD* = 10 years). For their graduate training in speech-language pathology, 71% of the SLPs attended a graduate institution that awarded research doctoral degrees and 29% attended a graduate institution that did not award research doctoral degrees. Classifications of graduate institutions (research versus non-research doctoral institution) were determined by the Carnegie Classification of Institutions of Higher Education (2015).

Evaluators. Evaluators in the archived data set included five certified SLPs from the same district as the 111 SLP participants. One of the five evaluators was the lead speech-language pathologist in the district, and the author of this dissertation. All evaluators described themselves as female, Caucasian, state licensed, and certified from the American Speech-Language-Hearing Association. Evaluators had a range of professional experience from 10 to 23 years (*M* = 18 years; *SD* = 6 years) and were randomly assigned to the 111 SLPs as their evaluators. Table 1 describes evaluator demographics.

Materials

Performance evaluation instrument. The performance evaluation instrument used in the SLP evaluations is found in Appendix A. This performance evaluation instrument was part of a formative performance evaluation process for the SLPs, designed to meet state educator accountability requirements and the district's continuous-improvement standards. The subsequent paragraphs briefly describe the development of the instrument per information supplied by the district of the SLPs. It is important to note the instrument was developed by a team of SLPs and this dissertation does not examine the methods in which the instrument was developed; rather, this dissertation examines the validity and reliability of this a pre-developed instrument.

The general timeline of instrument development was described by the district as follows: The SLP performance evaluation instrument was initially drafted in Fall 2013; revised in Winter 2014; piloted and further refined in Spring 2014; and fully implemented in the 2014-2015 school year.

More specifically, in Fall 2013, the overall design goal for the performance evaluation instrument was to create a *generic* performance evaluation instrument for district SLPs, akin to a generic performance evaluation instrument for district classroom teachers. The SLP performance evaluation instrument was intended to be used with

any SLP serving any age of student or speech-language disorder. The SLP performance evaluation instrument was modified from the district's 2013-2014 classroom teacher performance evaluation instrument, which was loosely based on the commonly-used Framework for Teaching (Danielson, 2007) and contained 31 instructional behaviors organized into four domains of practice: Planning and Preparation (seven items), Classroom Environment (five items), Instruction (nine items), and Professional Responsibilities (ten items).

During instrument development in Fall 2013, district teacher evaluation administration requested the SLP performance evaluation instrument contain approximately the same number of performance items (31) as the classroom teacher evaluation instrument. Where applicable, exact use of teacher evaluation items was encouraged; where not applicable, teacher evaluation items could be modified or completely deleted from the SLP instrument. To determine if the 31 teacher evaluation items could be applied to, modified for, or deleted from the SLP instrument, a five-member tool development team of district SLPs reviewed each teacher evaluation item using the following guiding question: Can this 2013-2014 classroom teacher evaluation item be observed for any SLP serving

- any grade/age of student (Birth to 22) …
- presenting any form of cultural/linguistic diversity and…
- experiencing any type of communication disorder (Feeding/Swallowing, Articulation, Language, Fluency, and/or Voice disorder) at…
- any severity level, for whom speech-language services may be delivered in …
- any location (home, classroom, and/or speech room) in …
- any service configuration (individual or group service)?

When the response was "no" for any of the six variables above, that particular classroom teacher item was modified, or not used, on the performance evaluation

instrument for SLPs. Sources of SLP item modification included standards of practice for school-based SLPs (ASHA, 2010); effective practices identified by SLPs (Ebert & Kohnert, 2010) and special educators (Jones & Brownell, 2014); and components of the evidence-based practice process for SLPs (ASHA, 2004).

In Winter 2014, district SLP focus groups were used to refine the SLP performance evaluation instrument and in Spring 2014, the instrument was piloted with seven SLPs and further edited. Finally, in May 2014, the district's speech-language pathology staff voted to implement the performance evaluation instrument found in Appendix A. This instrument was not altered during the 2014-2015 school year and contained 27 scaled behaviors (called elements) organized into four domains:

- Domain 1: Planning of Service (6 elements)
- Domain 2: Climate of Service (4 elements)
- Domain 3: Implementation of Service (6 elements)
- Domain 4: Professional Responsibilities, Due Process Documentation, and Case Management (11 elements)

For each of the 27 elements, performance indicators were used to position SLPs along a continuum of performance (*Requires Attention, Developing, Proficient,* and *Exemplary*). Examples of evidence for each continuum level were provided with key indicators (called "Look Fors").

Collaborating staff survey. One of the elements (Domain 4, Element 4) on the SLP performance evaluation instrument included gathering feedback on an SLP's ability to collaborate with staff members and other SLPs at their schools/sites. This feedback was gathered through an anonymous survey (Appendix B) in which all ratings were averaged across individuals invited to take the survey. To determine performance ratings for this element, SLPs submitted the names of five educational team members to the survey administrator. The five educational team members were not SLPs but rather

general and special education teachers, classroom assistants, school social workers, occupational and/or physical therapists, psychologists, nurses, and interpreters. Additionally, all SLPs serving the same school/site were invited to complete the survey for the SLP. Participants were provided the surveys electronically, with a submission deadline of two weeks after the initial email invitation. A total of 440 collaborating team members responded to the survey (response rate of 82%) and 193 collaborating SLPs responded to the survey (response rate of 97%).

In the Collaborating Staff Survey, participants were asked to rate the extent to which they agreed (*Strongly Disagree, Disagree, Agree,* or *Strongly Agree*) with statements describing the SLP's collaboration skills. For each survey item, the agreement ratings were assigned a numeric score of 1.0, 2.0, 3.0, 4.0, respectively; averaged across all survey respondents; and converted to a performance rating of *Requires Attention* (0.01 - .25), *Developing* (0.26 - 2.75), *Proficient* (2.76 - 3.75), or *Exemplary* (3.76 - 4.0).

Procedures

Evaluation process for SLPs. The evaluation process for SLPs consisted of three main contact points (pre-observation conference, observation, post-observation conference) with several intermediate steps. Performance ratings for the 27 elements were embedded in a multi-step process in which the majority of steps were predetermined by state regulations, district teacher evaluation administration, and the local teacher contract. Table 2 summarizes the steps of completion for SLPs, while the subsequent paragraphs briefly describe the evaluation process.

The SLP began the evaluation process by scheduling a 90-minute block of time with his/her evaluator on a randomly assigned date. This block of time included an observation of service to students that ranged from 30 to 45 minutes in length. SLPs were allowed to choose the session to be observed and while there was minor variation

in length of observations (*M* = 31.4; *SD* = 4.1), all sessions were identified by the SLPs as being representative of one "lesson" for an individual student or group of students.

Next, the SLP completed preparation guides for the pre-observation conference. In this pre-observation conference, the evaluator collected evidence for elements in Domains 1 (Planning of Service) and 4 (Professional Responsibilities, Due Process Documentation, and Case Management). Table 3 contains the pre-observation questions completed by SLPs. Appendix C contains an example of completed Domain 1 and 4 questions by an SLP in the 2014-2015 school year.

Immediately following the pre-observation conference, the SLP was observed for a student session that was self-selected by the SLP. While observing the SLP's session, the evaluator collected evidence for elements in Domains 2 (Climate of Service) and 3 (Implementation of Service). After the SLP observation, the evaluator sorted the evidence and assigned draft ratings for all 27 elements on the SLP performance evaluation instrument.

Independently and after the SLP observation, the SLP self-assigned draft ratings for the 27 elements. Additionally, the SLP self-described his/her observed student session using six variables: (a) grade of student (early childhood, primary, or secondary); (b) linguistic status of student (monolingual or multilingual for which languages); (c) communication disorder addressed in lesson (Language Disorder, Articulation Disorder, Fluency Disorder, Voice Disorder, or Combination); (d) perceived severity of student's communication disorder (mild, moderate, severe); (e) location of session (home, speech room, or classroom); and (f) group status of session (individual or group service).

Approximately one week after the SLP observation, a post-observation conference was scheduled. The SLP prepared for this post-observation conference by reading the evaluator's draft ratings and completing reflection questions. Post-

observation reflection questions are listed in Table 4. Appendix D contains an example of completed post-reflection questions for the same SLP in Appendix C. At the post-observation conference, the SLP and evaluator finalized performance ratings, based on consensus, and discussed areas of strength and professional growth. Finally, the SLP formally accepted or rejected the consensus ratings by signing off on the evaluation.

In summary, ratings for the 27 elements on the SLP performance evaluation instrument were determined by evidence collected by a speech-language pathology evaluator in a discussion of portfolio artifacts (Domains 1 and 4) and direct observation of service to students (Domains 2 and 3). Ratings for Domains 1, 2, and 3 were explicitly linked to the observation of student service, while ratings for Domain 4 were not necessarily linked to the observation of student service, as this domain reflected workload-wide, professional responsibilities. Table 5 summarizes the assessments of performance for each of the four domains. The time commitment to complete the entire evaluation process was estimated at 4-6 hours per SLP and 4-6 hours per evaluator. This was consistent with the time commitment for classroom teachers in their 2013-2014 performance evaluations.

At the end of the evaluation process, each SLP had three sets of ratings for the 27 elements on the experimental SLP performance evaluation instrument: (a) evaluator draft ratings, (b) SLP self-ratings, and (c) evaluator/SLP final, consensus ratings.

Training program for evaluators. SLP evaluators completed a six-session training program, totaling 30 hours in length, prior to evaluating the 111 SLPs. This training series was organized by the fifth evaluator (master coder) and was intended to provide evaluators experience in rating SLPs serving a wide range of ages and speech-language disorders. The focus on evaluator training was rating elements from Domains 1-3 and two elements (E7, E8) in Domain 4. Evaluators were not trained to rate the other nine elements (E1-E6, E9-11) in Domain 4. The rationale for this missing training

was because focus group feedback during tool development suggested indicators for Domain 4 were far more clear than indicators for Domains 1-3. Given time constraints for evaluator training, the master coder prioritized evaluator training for Domains 1-3 with limited training for Domain 4.

The first training session included orientating the evaluators to the performance evaluation instrument and indicators for each element. Next, a training video from an elementary SLP serving a student with Language Disorder was viewed and as a whole-group activity, the evaluators rated the SLP in the video on Domains 1-3 in the performance evaluation instrument. Ratings for each element were discussed until consensus was reached. During the second session of training, a training video from a different elementary SLP serving Articulation Disorder was viewed and again, as a whole-group activity, the evaluators rated Domains 1-3 and ratings were discussed until consensus was reached. During the third session of training, a training video from a secondary SLP serving Fluency Disorder was viewed, but rather than a whole-group activity, the evaluators independently evaluated the SLP for Domains 1-3. Ratings for each element were reviewed, compared for agreement, and discussed when there were differences in ratings. Procedures from the third session of training were repeated for additional fourth training session, using a video from a preschool SLP serving combined Language Disorder and Articulation Disorder. Finally, a fifth training video from an elementary SLP serving a student using an Augmentative and Alternative Communication device was viewed and independently rated for Domains 1-3. During this final video, evaluators were at least 80% perfect agreement with each other for each element in Domains 1-3.

After the five training sessions above, the sixth training session focused on rating Elements 7 and 8 from Domain 4. Five de-identified speech-language evaluation reports and individual educational plans were read and rated by the evaluators. Following the

same procedures for the training videos, percent agreement was calculated on the fifth and final evaluation report and individual educational plan. During this final rating, evaluators were 100% and 80% perfect agreement for Elements 7 and 8, respectively.

Procedures for gathering face validity data. After the evaluation process, the 111 SLPs were invited to complete two optional surveys regarding their evaluation of the performance evaluation instrument and their evaluators. Both surveys were developed by the district's teacher evaluation department and the surveys completed by SLPs closely resembled the surveys completed by classroom teachers. SLPs were provided the surveys electronically and those who completed the surveys did so anonymously. The deadline to complete both surveys was two weeks after the initial email invitation.

In the first survey (Appendix E) participants were asked to rate the extent to which they agreed (*Strongly Disagree, Disagree, Agree,* or *Strongly Agree*) the 27 elements represented effective speech-language pathology practices and the indicators appropriately placed SLPs on a continuum of performance. Participants were able to provide optional comments related to each element and indicators. Of the 111 SLPs, 47 responded to the survey (response rate 42.3%).

In a separate survey (Appendix F) participants were asked to rate the extent to which they agreed (*Strongly Disagree, Disagree, Agree,* or *Strongly Agree*), with optional comments, their evaluator provided clear evidence of performance aligned to the evaluation instrument. Of the 111 SLPs, 73 responded to the survey (response rate 65.8%) of five questions.

Procedures for gathering interrater reliability data. Interrater agreement is the extent of agreement between two or more raters. To estimate interrater reliability, 34 of the evaluation sessions were randomly selected and video recorded with consent of the SLP and family of the student observed. This number represented 30% of all evaluations (34/111). Once recorded, the videos were uploaded to a secure district

server accessible only to the evaluators. All videos were rated by May 31, 2015. Although the 34 interrater videos were randomly selected from all five evaluators, the fifth evaluator did not participate in interrater reliability coding due to scheduling constraints that prevented rating by May 31, 2015; only evaluators one, two, three, and four participated in interrater reliability coding. Once the videos were uploaded and accessible, the four evaluators independently rated the 27 elements on the performance evaluation instrument and the videos were permanently deleted one week after the coding. When rating the videotaped sessions, evaluators were instructed not to replay the videos unless there was poor audio feed or technical difficulties due to the video camera. These instructions were intended to approximate the "real time" evaluation process. To ensure consistently in coding interrater reliability videos, the evaluators followed the protocol in Appendix G.

Analytic Strategy

The archived data file used in this study contained six spreadsheets of de-identified data: (a) evaluator draft ratings for 111 SLPs, (b) self ratings for 111 SLPs, (c) evaluator/SLP final, consensus ratings for 111 SLPs, (d) interrater reliability ratings from four evaluators for 34 videotaped sessions, (e) instrument survey feedback from 47 SLPs, and (f) evaluator survey feedback from 73 SLPs. Data in these spreadsheets included ordered, categorical variables measured at a static point in time. In the first four spreadsheets, the ordered, categorical data included performance ratings based on a continuum of performance (*Requires Attention, Developing, Proficient,* and *Exemplary*). In the last two spreadsheets, the ordered, categorical data included survey opinion ratings based on a continuum of agreement (*Strongly Disagree, Disagree, Agree,* and *Strongly Agree*). To address the research aims in this study, data from these six spreadsheets were described and analyzed. The analytic strategy for each research aim is below.

Research Aim 1. To describe the performance of educational SLPs on the experimental performance evaluation instrument, two sets of analyses were run on data from three of the six spreadsheets in the archived data file: (a) evaluator draft ratings for 111 SLPs, (b) self-ratings for 111 SLPs, and (c) evaluator/SLP final, consensus ratings for 111 SLPs.

The first set of analyses included descriptive analyses summarizing the percentage of occurrence for the four performance levels (*Requires Attention, Developing, Proficient,* and *Exemplary*) across each of the 27 elements on the evaluator draft ratings and SLP self-ratings. Additionally, a descriptive analysis of the frequency count and direction of change from the evaluator draft to the evaluator/SLP consensus ratings was completed across each of the instrument's 27 elements. A test of differences in means (paired-samples t test) was used to determine whether evaluators scored SLPs differently than SLPs themselves in each of the four instrument domains. To compare means, a domain score was computed for each SLP by summing up the ratings for all instrument elements within the domain. In this computation, *Requires Attention*, *Developing*, *Proficient*, and *Exemplary* were coded with values of 1.0, 2.0, 3.0, and 4.0.

The second set of analyses involved examining the possibility that SLP and/or student/session characteristics impacted evaluator ratings. If SLP and/or student/session characteristics impact evaluator performance evaluation ratings, bias may be present in the experimental SLP performance evaluation instrument (Park, Chen, & Holtzman, 2014; Holdheide et al., 2010). The potential relationship between performance evaluation ratings and SLP and student/session characteristics was examined through a linear mixed effects (LME) model. A LME model allows a researcher to express a relationship in data as a function, in order to determine the extent to which a dependent variable may be predicted from one or more independent

variables (Winter, 2013). In this study, an SLP's total score from his/her evaluator draft ratings was modeled (predicted) as a function of the SLP and student/session characteristics. The total score for an SLP was computed by summing up the evaluator draft ratings for each of the 27 instrument elements. In this computation, *Requires Attention*, *Developing*, *Proficient*, and *Exemplary* were coded with values of 1.0, 2.0, 3.0, and 4.0.

The "mixed" part of LME refers to a model that includes both "fixed" and "random" effects as independent (predictor) variables. Fixed effects have levels that do not vary across individuals and are exhaustive in the data set; for example, the student grade session/student variable in this study is a fixed-effect predictor because there are only three potential grade levels for any student served by an SLP. Random effects have many possible levels that could vary across individuals and potentially, these variations could impact the dependent variable. In this study, evaluator was considered a random effect in the LME model, as there may be variations in performance ratings between evaluators and/or correlations in ratings from the same evaluator. In order to predict a dependent variable from any fixed effects, random effects must be controlled for in an LME model (Winter, 2013).

The statistical program R (R Core Team, 2012) running the analysis package *lme4* (Bates, Maechler, Bolker, & Walker, 2014) was used to perform the LME analysis of the relationship between evaluator draft ratings and SLP and student/session characteristics, after controlling for random effects of evaluator. Prior to fitting the LME model, assumptions of normality were tested to make sure an LME model was the appropriate statistical model for the data under study.

The following fixed effects (without interaction terms) were entered into the LME model: (a) years of experience as an educational SLP, which was a continuous variable; (b) type of SLP graduate training institution, which contained two levels for research

institution or non-research institution; (c) grade of student observed, which contained three levels for early childhood, primary, or secondary grades; (d) linguistic status of student, which contained two levels for monolingual English or multilingual; (e) communication disorder addressed in lesson, which contained four levels for Language Disorder, Articulation Disorder, Fluency Disorder, or Combination of Language and Articulation Disorder; (f) perceived severity of student's communication disorder, which contained three levels for mild, moderate, or severe disorder; (g) location of session, which contained three levels for home, speech room, or classroom locations; and (i) group status of session, which contained two levels for individual or group service. Length of observation, a continuous variable, was also a fixed effect, in the event the length of observation impacted evaluator ratings. The LME model was fit as follows:

SLP Total Score ~ SLP experience + SLP training + Student Grade + Student Linguistic Status + Student Disorder Type + Student Disorder Severity + Session Location + Session Group Status + Session Length + (1|evaluator)

In terms of interpreting LME model findings, impacts of the fixed effects were estimated from the LME coefficients table. This table contained the intercept coefficient for the LME model, which was the predicted value of the dependent variable (SLP total score) when all independent variables were valued at 0. In addition to the intercept value, the table contained the coefficients that represented the changes in predicted SLP total score value due to each individual fixed effect, after holding all other independent variables constant. Each fixed effect change was associated with a standard error, degrees of freedom, t-value, and *p*-value. Impact of the random effect was estimated from the LME variance table, which listed the amount of variance (and standard deviation) in the predicted score value due to evaluators. To determine if this random effect variance was significantly greater than zero, a Chi-squared test was conducted, with a *p*-value as an indicator of statistical significance.

Research Aim 2. To determine the extent to which items on the experimental SLP evaluation instrument demonstrated preliminary construct and face validity several types of analyses were run on data from three out the six spreadsheets in the archived data file: (a) evaluator draft ratings for 111 SLPs; (b) instrument survey feedback from 47 SLPs; and (c) evaluator survey feedback from 73 SLPs. The first spreadsheet was used to examine evidence for construct validity while the latter two spreadsheets were used to examine evidence for face validity.

***Construct validity*.** Construct validity is defined as the degree to which a test measures what it intends to measure (Kimberlin & Winterstein, 2008). The intent of the experimental SLP performance evaluation instrument was to measure effective speech-language pathology practices in four domains. Elements were developed within each domain based on the assumption those elements were measuring a cohesive construct of effective speech-language pathology practice. For example, the six elements in Domain 1 were developed as an indication of an SLP's skill in planning effective intervention for students. Determining if elements within a domain represent an intended construct has practical implications for using the performance evaluation data. If elements within a domain are related to each other, this provides evidence they are measuring similar information about a speech-language pathology practice and ratings for individual elements could be averaged to summarize an SLP's performance for that domain. If elements within a domain are not related to each other, they may not be measuring the same construct and likely should not be averaged to indicate overall domain performance.

In this study, two sets of analyses were used to examine preliminary construct validity. I am using the term "preliminary" as a descriptor here because the analyses represented a first attempt at providing construct validity evidence for an initial-version performance evaluation instrument. An accumulation of evidence, using multiple

statistical methods, is required to establish solid construct validity for an instrument (Kimberlin & Winterstein, 2008).

The first set of construct validity analyses included item-item and item-total score correlations on the evaluator draft ratings within each domain. Item-item correlations indicate the degree and direction of relationship between instrument items while item-total correlations indicate the degree and direction of relationship between an individual item and an SLP's total score. The total score was computed for each SLP by summing up the ratings for all 27 instrument elements. In these computations, *Requires Attention*, *Developing*, *Proficient*, and *Exemplary* were coded with values of 1.0, 2.0, 3.0, and 4.0. Bivariate Spearman correlation coefficients were used in the correlational analyses, with *p*-values indicating statistical significance of the correlations.

The second set of construct validity analyses included conducting an exploratory factor analysis (EFA) on evaluator draft ratings using IBM SPSS Statistics 24.0 (IBM SPSS Statistics, 2016). A factor analysis is a statistical technique to analyze potential relationships between variables and explain how related variables may group into dimensions/factors. The EFA was used to help to answer the following questions: Were SLPs' skills on the experimental performance instrument indicative of one, large professional competency factor, or several smaller professional competency factors?; and, if SLP skills grouped into several smaller factors, what were the factors, were they consistent with the four a priori domains, and which instrument elements best represented those factors? Prior to conducting the EFA, assumptions of factorability were verified to determine if the data were suitable for EFA.

When conducting an EFA, there are two main results from the analysis. The first result pertains to how well a set of factors explains the variance in the data; the second result pertains to the factor loadings for a model. A good factor model explains the most amount of variance with the least number of factors. An eigenvalue is the amount of

total variance explained by each factor. In an eigenvalue results table, each factor has an eigenvalue and that value is converted to a percentage of overall variance accounted for by that factor. Cumulatively, a set of factors should explain approximately 50% or more of the overall variance and factors that do not appreciably add to the cumulative explained variance should be deleted (Stevens, 1992). Typically, factors with eigenvalues below 1.0 do not appreciably add the cumulative explained variance based on Kaiser's Rule (Kaiser, 1960). A scree plot is a visual display of eigenvalues and is an easy way to observe how many factors are explaining the majority of overall variance in data.

In an EFA, a researcher first enters all potential factors in a full factor model and based on eigenvalues, determines if that full model best explains the variance in the data, or, if a model with fewer factors better explains the variance in the data. In this study, the full-factor model included 27 factors, each factor representing an element on the experimental SLP performance evaluation instrument. The resulting eigenvalue table and scree plot demonstrated only a handful of factors explained the majority of overall variance in evaluator's draft ratings.

After determining the best factor model from the eigenvalues and scree plot, the degree and direction of relationship between each factor and individual variable were determined. This relationship is called a factor loading and represents the correlation between each variable and their factor. Factor loadings of equal to or greater than .50 suggest the variable is a good representative of the factor; .30 and .50 suggest the variable may be a fair representative of the factor; and below .30 suggests the variable is not representative of the factor (Stevens, 1992).

***Face validity*.** In addition to construct validity, a second type of validity, face validity, was examined in this study. Face validity is the degree to which stakeholders believe an assessment accurately measures what it claims to measure (Kimberlin &

Winterstein, 2008). In this study, face validity was estimated using survey feedback data from the SLPs within the district. Two analyses were used to gauge preliminary face validity.

The first analysis included a descriptive analysis summarizing the percentage of occurrence for the four levels of agreement (*Strongly Agree, Agree, Disagree,* and *Strongly Disagree*) in SLPs' evaluation of the performance evaluation instrument elements and indicators. The second analysis included a descriptive analysis summarizing the percentage of occurrence for the four levels of agreement (*Strongly Agree, Agree, Disagree,* and *Strongly Disagree*) in SLPs' evaluation of their evaluators.

Research Aim 3. To determine the extent to which items on an experimental SLP evaluation instrument demonstrated agreement and produced reliable results, a generalizability study (G-study) was conducted on data from the archived data spreadsheet containing interrater reliability ratings from the four evaluators across 34 videotaped sessions. In total, 3672 interrater ratings were entered into a G-study (34 SLPs x 27 instrument items x 4 raters = 3672 ratings). The software program EduG (Swiss Society for Research in Education Working Group, 2012) was used to conduct the G-study.

A generalizability study (G-study) is a type of reliability analysis that allows an investigator to examine multiple sources of variance (called facets) in a single planned analysis and can uncover variance issues not otherwise detected by traditional interrater agreement measures such as kappa (Hill et al., 2012a). G-studies partition variance into component parts, their interactions, and residual measurement error. In contrast, traditional kappa measures for interrater reliability evaluate one source of variance (the rater) and do not attend to other sources of variance, for example, variance that may due to instrument items, session/lessons, or data collections methods. Generalizability theory

is commonly used in performance evaluation research (e.g., Hill et al., Ho & Kane, 2013; Semmelroth & Johnson, 2014; Lawson, 2015).

The advantage of a G-study over traditional kappa measures is illustrated in Hill et al. (2012a), who explain how it is possible to make incorrect inferences about the reliability of educator performance evaluation instruments based on the results of kappa measures. For example, in a traditional kappa measure, it is possible to have 98% interrater agreement for a specific instrument item because the item was not actually observed frequently. In this situation, a high agreement rating of 98% might falsely suggest an item was reliably observed, when in fact, it was hardly present. In a G-study, however, items that did not occur frequently account for very little variance in the instrument's overall variability and thus, may be excluded from subsequent iterations of an instrument development because these items do not adequately distinguish performance levels among individuals.

There are two main results generated by a G-study. The first result includes the decomposition (percentage) of variance explained by each source of variance, their interactions, and the residual error. For each of the four domains, the G-study in this study included three sources of potential variance: raters, items, and SLPs. Table 6 describes the type variance components resulting from this three-facet design. In terms of interpretation, variance due to raters is ideally low, indicating raters did not differ in scoring the same SLP and conversely, variance due to SLPs is ideally high, as their performance represents what is being studied and the developers of the performance evaluation instrument intended to differentiate performance among SLPs.

The second result generated in a G-study is an overall indication of reliability, called the G-coefficient, which is interpreted analogous to classical reliability coefficients. There is variation in what is an acceptable G-coefficient, but .65 has been cited as acceptable in educator performance evaluation research with .80 cited as a traditional

cutoff level (Ho & Kane, 2013). There are two types of G-coefficients, relative and absolute. Relative G-coefficients are often used in ranking decisions, in which the reliability of the instrument is used when determining an SLP's performance relative to other colleagues. Absolute G-coefficients are often used in criterion/mastery decisions, in which the reliability of the instrument is used to understand how well an SLP performed against a criterion/mastery level. In practice, relative coefficients are more common in "low-stakes" personnel decisions while absolute coefficients are more common in "high-stakes" personnel decisions (e.g., promoting or demoting staff based on performance evaluation ratings) (Lawson, 2015).

Finally, given its ability to partition sources of variability, results from a G-study can be used to conduct a follow-up decision study (D-study). A D-study is used to identify optimal future study design based on a desired G-coefficient (e.g., increasing in the number of raters to reach a .65 coefficient). A D-study results table contains relative and absolute G-coefficients for different study conditions. In this study, a D-study was conducted for each domain to determine the relative and absolute G-coefficients under the conditions of one, two, three, or four raters.

Summary

This chapter described the research design and analytic strategy for examining preliminary validity and reliability evidence for an experimental performance evaluation instrument for educational SLPs. The study used archived performance evaluation data from the 2014-2015 school year from 111 SLPs and 5 evaluators within a single mid-sized urban district to:

1. To describe the performance of SLPs on an experimental evaluation instrument, using descriptive data analysis of frequency counts and a linear mixed model of effects.

2. To determine the extent to which items on an experimental SLP evaluation instrument demonstrated construct and face validity, using descriptive data analyses, an exploratory factor analysis, and survey methods.
3. To determine the extent to which items on an experimental SLP evaluation instrument demonstrated agreement and produced reliable results, using generalizability-study and decision-study analyses.

The next chapter describes the results of the descriptive and inferential analyses.

Chapter Four: Results

Research Aim 1

The first research aim in this study was to describe the performance of educational speech-language pathologists (SLPs) on an experimental performance evaluation instrument (see Appendix A). Figure 1 describes the distribution of evaluator draft and collaborating staff survey ratings (Domain 4, Element 4) across the 111 SLPs. The 27 instrument elements are organized by Planning of Service (Domain 1), Climate of Service (Domain 2), Implementation of Service (Domain 3), and Professional Responsibilities, Due Process Documentation, and Case Management (Domain 4). Collapsed across all elements, the average percentage of occurrence for evaluator draft ratings was 41% *Exemplary,* 49% *Proficient*, 9% *Developing*, and 1% *Requires Attention*.

Figure 2 describes the distribution of ratings when the 111 SLPs self-rated their performance. SLPs did not self-rate the collaborating staff survey (Domain 4, Element 4), so there are no data for D4, E4 in Figure 2. Collapsed across all elements, the average percentage of occurrence for the self-ratings was 23% *Exemplary,* 64% *Proficient*, 12% *Developing*, and 1% *Requires Attention*. On average, SLPs rated themselves lower in performance than evaluators in all four domains of practice. Paired t-tests showed a significant difference between evaluator draft ratings (M = 3.21; SD = 0.37) and self ratings (M = 3.03; SD = 0.40) for Domain 1 ($t(110) = 5.76$, $p < .001$); evaluator draft ratings (M = 3.60; SD = 0.36) and self ratings (M = 3.33; SD = 0.47) for Domain 2 ($t(110) = 6.55$, $p < .001$); evaluator draft ratings (M = 3.23; SD = 0.43) and self ratings (M = 3.06; SD = 0.44) for Domain 3 ($t(110) = 4.06$, $p < .001$); and evaluator draft ratings (M = 3.24; SD=0.33) and self ratings (M = 3.10; SD = 0.34) for Domain 4 ($t(110) = 4.18$, $p < .001$).

Table 7 summarizes the frequency of occurrence and direction of change from the evaluator draft to evaluator/SLP final consensus ratings for each of the instrument's

elements. Overall, 54 evaluator draft ratings were changed, representing 1.87% of all evaluator draft ratings. Of the 1.87% changes, 1.59% were changed in the positive direction (e.g., moving from *Developing* to *Proficient*) and 0.28% were changed in the negative direction (e.g., moving from *Proficient* to *Developing*). All changes included one scale level and not greater than one scale level; for example, no ratings were changed two scale levels such as moving from *Developing* to *Exemplary* or *Exemplary* to *Developing*. Of the 111 SLPs, the 54 changes were from 31 different SLPs. Of the five evaluators, 24 of the 54 changes fell under evaluator two, while 11, 10, seven, and two changes fell under evaluator one, five, four, and three, respectively.

Table 8 summarizes the frequency counts of the six student and session variables for the 111 observations. The frequency counts suggest the SLPs served a diverse group of students during their observations. Approximately half of the students were in primary school; just over 40% were bilingual; and observations for service of Language Impairment, or combined Language Impairment and Articulation Impairment, were selected by 75% of the SLPs. Almost 90% of the observations included students with perceived moderate or severe disabilities. Approximately 75% of the observations were conducted in the speech room and similarly, approximately 75% of the observations included service to individual students.

Table 9 presents the coefficients table from the linear mixed effects (LME) model that was used to examine the potential impact of SLP and student/session variables on evaluator draft ratings. On the coefficient table, the reference level for each variable was alphabetically arranged. The maximum total score an SLP could receive from evaluators was 108 points (27 elements x 4.0 points for each element). The LME model results showed two of the six student/session variables (group status of session and multilingual status of student) were not statistically significant predictors of an SLP's total score, while four out of the six student/session variables were statistically significant

predictors of SLP total score. The impact of the four significant student/session ranged from 6-10 points, after holding effects from all other variables constant. More specifically, the age range Birth-5 was associated with a 10-point drop in total score (compared to primary and secondary age ranges); combined language and articulation disorder was associated with a 6-point drop in total score (compared to articulation, language, or fluency disorder, only); mild disorder was associated with an 8-point increase in total score (compared to moderate and severe disorders); and the location *speech room* was associated with a 9-point drop in total score (compared to classroom and home locations). Length of session was not a significant predictor of SLP total score, nor was an SLP's training institution. There was a small, but significant effect of SLP years of experience on total score, in which every year of professional experience as an educational SLP was associated with a 0.2-point drop in total score, after holding all other variables constant. The Chi-squared test showed the type of evaluator (evaluator one, two, three, four, or five) was not a significant predictor of SLP total score ($X^2(1, N = 111) = 2.06$, $p = .20$).

Research Aim 2

The aim of the second research question was to determine the extent to which items on an experimental SLP evaluation instrument demonstrated preliminary construct and face validity.

Construct Validity. Tables 10-13 present the first set of construct validity analyses, including the item-item and item-total score correlations on evaluator draft ratings for each domain. Conventional interpretation values (Evans, 1996) can be used to indicate the strength of relationship for statistically significant correlation coefficients: 0-.19 = very weak relationship, .20-.39 = weak relationship, .40-.59 moderate relationship, .60-.79 = strong relationship, and .80-1.00 = very strong relationship. While there were many significant positive item-item correlations, few were

strong in strength. The three highest item-item correlations were for items in due process compliance paperwork (Domain 4, E7 and E8; .62 correlation), data collection and decision-making (Domain 4, E10 and E11; .59 correlation), and planning coherent, functional service (Domain 1, E2 and E3; .55 correlation).

Compared to item-item correlations, there were relatively more item-total correlations strong in strength, indicating some instrument items were more closely related to an SLP's total score than other items. Item-total coefficients above .60 included correlations between an SLP's total score and planning of evidence-based practice intervention (D1, E1) that placed learning targets in functional contexts (D1, E3); delivery of intervention that was cognitively engaging (D3, E4), monitored for student progress (D3, E5), and designed to promote student independence (D3, E6); and an SLP's evaluation reports (D4, E7), programming of dynamic service delivery (D4, E9), and data-based decision making (D4, E11).

As the second set of construct validity analyses, Figure 3 and Tables 14-15 describe the results from the exploratory factor analyses (EFA) on evaluator draft ratings. The EFA was used to examine if and how elements on the instrument grouped into factors, or underlying structures. Given the evaluator ratings were not normally distributed, the EFA was run using a principal axis factoring extraction method (Yong & Pearce, 2013). To maximize loadings onto a factor, a varimax with Kaiser normalization rotation method was used. Prior to running the EFA, assumptions of factorability were verified through a Kaiser-Meyer-Olkin (KMO) sampling adequacy statistic and Bartlett's Test of Sphericity. The KMO statistic was .812 (values of .80 – 1.0 are acceptable; Stevens, 1992) and the Bartlett's test was significant (X^2 = 1145.76, *df* = 351, $p < .001$). These test statistics indicated collinearity among the variables, making the data suitable for EFA.

Figure 3 is the scree plot of eigenvalues for the full factor model, in which all 27 elements on the instrument were entered as components (factors) in the EFA. From this figure, one can observe eight or less components explained the majority of overall variance in evaluator draft ratings. Table 14 contains the eigenvalues and percentage of variance explained by each component. From this table, four factors explained approximately 50% of the variance, meeting the minimum requirements of explained variance for a preliminary EFA (Stevens, 1992). Factor loadings for the four-factor model are presented in Table 15. Each factor loading represents the correlation the element and factor.

In general, there were several factor loadings above .30 in Table 15, suggesting those elements were a fair representative of the factor (Stevens, 1992). There were two elements, D4, E1 (use of self-assessment and self-reflection) and D4, E4 (collaborating staff survey) that did not appreciably load onto any factor. While four factors emerged in the EFA, these factors were only somewhat consistent with the four a priori domains in the performance evaluation instrument. That is, not all Domain 1 elements clustered into one factor, nor did all elements in Domains 2, 3, and 4 elements cluster into respective factors. Rather than strictly clustering into a priori domains, four of six Domain 1 elements clustered into a factor (Factor #3 in Table 15), the majority of Domain 2 and 3 elements clustered into another factor (Factor #1 in Table 15), and only a few Domain 4 elements clustered into a third factor (Factor #4 in Table 15). A fourth factor (Factor #2 in Table 15) contained a combination of elements from Domains 1, 2, 3, and 4.

Given the instrument elements did not fully cluster in their four a priori domains, I will attempt to name an alternative structure for the observed factor loadings. The first factor might be called “Service Planning” (Factor #3 in Table 15) and includes the planning of intervention/instruction that is culturally relevant and personally meaningful (D1, E4); functional (D1, E3); connected to prior learning (D1, E2); implemented in the

least restrictive environment (D4, E9); and consistent with a family's priorities, hopes/wishes, and concerns for their child (D4, E5). A second factor might include "Service Implementation" (Factor #1 in Table 15) and includes the implementation of intervention/instruction that is engaging (D3, E2), organized (D2, E1; D2, E2; D2, E3; D1, E5), based on high expectations for students (D3, E3; D3, E4), promotes positive relationships with students (D2, E4); and provides quality feedback to students about their performance (D3, E6). A third factor might include "EBP Processes for Service" (Factor #2 in Table 15) and includes elements that focus on evidence-based practice planning (D1, E1), selection and communication of learning targets to students (D3, E1), and progress monitoring of and data-based adjustments for intervention (D1, E6; D3, E5; D4, E10; D4, E11). Additionally, a SLP's engagement in and implementation of professional development training (D4, E3) was associated with their EBP processes. Finally, a fourth factor might include "Compliance with Legal Mandates" and includes the completion of evaluations (D4, E7), educational plans (D4, E8), and response-to-intervention, pre-referral procedures (D4, E6). An SLP's data-based decision making (D4, E11) was associated with compliance of these legal mandates.

Face Validity. In addition to construct validity, face validity was examined in this study, as estimated from survey feedback from the SLPs' evaluation of the performance evaluation instrument and evaluators. Tables 16-17 summarize the percentage of occurrence for each of the four agreement levels (*Strongly Disagree, Disagree, Agree, Strongly Agree)* in SLPs' evaluation of instrument elements and indicators. Optional participant comments from this survey are reported in Appendix H, Table H1. Of 111 SLPs participating in the study, 42% responded to the feedback survey. These responders generally indicated the instrument's 27 elements represented effective speech-language pathology practices; the combined percentage of *Strongly Agree* and

Agree ratings was 90% or above for any given element, with an average of 97% combined *Strongly Agree* and *Agree* ratings across elements.

However, not all of the 42% of survey respondents indicated the indicators appropriately placed SLPs on a continuum of performance; here, the combined percentage of *Strongly Agree* and *Agree* ratings was 80% or above for any given indicator, with an average of 93% combined *Strongly Agree* and *Agree* ratings for indicators. The highest percentage of disagreement ratings for indicators were for the following elements: Designs speech/language sessions that place learning targets in functional contexts (D1, E3), Plans for assessment strategies to monitor student progress (D1, E6), and Explicitly communicates speech/language learning objectives (D3, E1). Optional comments suggest the wording for these indicators was not necessarily clear and/or expectations for *Exemplary* were not necessarily attainable for all Birth-22 SLPs. Across the 27 elements, there were more optional comments provided on indicators (*n* = 56) compared to elements (*n* = 17).

Table 18 summarizes the percentage of occurrence for each of the four agreement levels (*Strongly Disagree, Disagree, Agree, Strongly Agree)* in SLPs' evaluation of their evaluators. In general, the 73 SLPs who responded to the survey felt favorably about their evaluator's ability to provide specific feedback and clear evidence of performance ratings. The combined percentage of *Strongly Agree* and *Agree* ratings for any given question was 96% or above.

Thirty-seven of the 73 respondents provided optional comments in the survey (Appendix H, Table H2). The number of optional comments provided for each evaluator was similar across evaluators; evaluator one, two, three, four, and five received ten, eight, five, seven, and seven comments, respectively. The agreements ratings from the 73 respondents were consistent with the optional comments. Thirty-six comments suggested the evaluators provided specific feedback that was perceived as helpful or

valuable to the SLPs while one comment (29) suggested the evaluation process was not particularly helpful or valuable. Three comments suggested there may be room for improvement by decreasing the time it takes to complete the evaluation process (9, 17) and increasing the opportunities for coaching by the evaluators (28).

Research Aim 3

The aim of the third research question was to examine evidence of reliability for elements on the experimental SLP performance evaluation instrument. To address this aim, a generalizability study (G-study) was conducted to decompose sources of variance and estimate reliability in the ratings from four evaluators across 34 videotaped sessions. Student and session characteristics of the 34 sessions are presented in Table 19. In general, the distribution of student and session characteristics of the 34 interrater videos was similar to the distribution of student and session characteristics of the entire 111 observation sessions (Table 8).

Results of the G-study were organized according to the four domains in the instrument: Planning of Service (Domain 1); Climate of Service (Domain 2); Implementation of Service (Domain 3); and Professional Responsibilities, Due Process Documentation, and Case Management (Domain 4). Table 20 reports the variance decomposition for each domain, including the percentage of variance explained by SLPs (s); raters (r); items (i); the interaction of SLPs, raters, and items (s * r; s * i; r * i); and the highest-order interaction effect (s * r * i), confounded with residual error (e).

Across the instrument domains, results showed 5%-27% of the total variance in performance ratings was attributable to differences in SLPs. Percent variance due to SLPs was highest for Domains 1 (27%) and 3 (21%) and lowest for Domain 2 (5%). Percent variance due to raters was minimal (less than 5% across domains), but when observed, rater variance occurred in the interaction of SLP x Rater, indicating that some raters scored some SLPs higher than others. Item variance was also present,

particularly for Domain 2, in which the type of instrument item accounted for the majority (64%) of the variance in Domain 2 ratings. Finally, the highest interaction variance confounded with residual error, accounted for 13%-31% of the total variance, indicating there was a considerable amount of unmeasured variance for all four domains.

Absolute and relative G-coefficients are indicators of overall reliability (Hill et al., 2012a). When absolute coefficients are calculated, all sources of variance contribute to measurement error, with the exception of the object under study; for this study, those components included the variance due to raters (r); items (i); the interaction of SLPs, raters, and items (s * r; s * i; r * i); and the highest-order interaction effect (s * r * i), confounded with residual error (*e*). When relative coefficients are calculated, only those sources of variance that interact with the object of measure contribute to measurement error; for this study, those variance components included the interaction of SLPs and raters (s * r); SLPs and items (s * i); and the highest-order interaction effect (s * r * i), confounded with residual error (*e*).

Under the original interrater conditions of four raters per videotaped session, results showed absolute G-coefficients of .74, .19, .69, and .71 and relative G-coefficients of .80, .48, .73, and .74, for Domains 1, 2, 3, and 4, respectively. Using .65 as an acceptable G-coefficient (Ho & Kane, 2013), these results suggested acceptable reliability for Domains 1, 3, and 4 using four raters when making absolute or relative decisions. When interpreted analogous to classic measurement theory (.80 as the traditional cutoff level), only Domain 1 demonstrated adequate reliability under the condition of four raters in a relative decision.

Having four raters per SLP is likely impractical for any school setting and thus, the decision study (D-study) was an important final step in the G-study analyses. Here, I asked the question: Can an acceptable reliability of .65 be obtained with *fewer* than four raters? Table 21 reports the changes in absolute and relative G-coefficients (with

standard error of measurement) under the conditions of one, two, three, or four raters. With just one rater, an acceptable reliability level of .65 was obtained for Domains 1 and 4. With two raters, an acceptable .65 level was obtained for Domain 3. Even with four raters, however, an acceptable reliability .65 level could not be obtained for Domain 2. Across all domains, adding a second rater increased the absolute or relative G-coefficient and decreased the SEM. However, having *more* than two raters did not appreciably increase G-coefficients or decrease the SEMs for Domains 1 and 4 and only somewhat increased the G-coefficients and decreased the SEMs for Domains 2 and 3. In sum, the D-study showed minimal reliability benefits of employing more than two raters.

Chapter Five: Discussion

This study was designed to examine validity and reliability evidence for an experimental performance evaluation instrument for school-based speech-language pathologists (SLPs). Three research aims were addressed in the study. Results from the current study are reviewed in relation to these three research aims as well existing educator performance evaluation research. Following this discussion, implications for practice, limitations of the study, and recommendations for further research are described.

Research Aim 1

Distribution of performance ratings. The first research aim included describing the performance of the 111 SLPs on the performance evaluation instrument. Results showed evaluators used a restricted range of categories when assigning performance ratings to SLPs; specifically, the distribution of evaluator draft ratings was 41% *Exemplary*, 49% *Proficient*, 9% *Developing*, and 1% *Requires Attention*. Rarely (less than 2% of the time), these draft ratings changed during evaluators' post-observation discussions with SLPs. This restricted range of SLP performance ratings was similar to the restricted range of performance ratings reported for classroom teachers. A recent meta-analysis of 19 states' classroom teacher performance evaluation data showed that, on average, evaluators used the ratings of *Exemplary*, *Proficient*, *Developing*, and *Unsatisfactory* 39%, 58%, 2%, and 1% of the time (Kraft & Gilmour, 2017). Additionally, SLPs self-rated their performance lower than evaluators, a phenomenon that has also been reported for some areas of teaching in classroom teacher observations (Gitomer et al., 2014). In Gitomer et al., teachers typically underestimated their classroom organizational skills but overestimated aspects of their instructional quality, compared to external evaluators.

There are several reasons that may explain the restricted range of performance ratings observed on the experimental SLP performance evaluation instrument. To begin, the scales for the instrument's items may not have been fine enough to distinguish performance among SLPs. Second, SLPs' performance on the instrument items may not have varied enough to make distinctions in performance and third, evaluators may not have felt comfortable enough making distinctions in performance on instrument items.

It is possible that bigger differences in performance ratings among the SLPs may have been found if there were bigger differences in speech-language graduate program accreditation and certification standards. Speech-language pathology graduate programs must meet minimal standards of professional education quality from their national accreditation agency (Council on Academic Accreditation in Audiology and Speech-Language Pathology, a semi-autonomous body of the American Speech-Language-Hearing Association, 2017). Additionally, speech-language pathology graduate programs must meet minimal standards of clinical education quality from their national certification agency (Council for Clinical Certification in Audiology and Speech-Language Pathology, a semi-autonomous body of the American Speech-Language-Hearing Association, 2013). To maintain certification, certified SLPs (96% of SLPs in this study) must engage in a cycle of continuing education credits. All SLPs in the current study completed a master's degree program, which is the entry degree for the field. If there was an alternative to a master's degree in order to practice speech-language pathology, there might have been larger differences in performance evaluation ratings. In a study of student achievement outcomes for students in special education, Feng and Sass (2010) showed achievement scores were higher for students when their teachers completed a traditional post-baccalaureate program versus an alternative (e.g., "fast-track") training program for special education teaching. The low frequency of

Requiring Attention and *Developing* ratings for SLPs in this study may have been due to the quality of standards in place for speech-language pathology graduate programs and certification maintenance.

Alternatively, or additionally, data from Kraft and Gilmour (2017) suggest a low frequency of *Requiring Attention* and *Developing* performance evaluation ratings may be due to conscience decisions by evaluators to avoid assigning these ratings. In their study of challenges evaluators face, Kraft and Gilmour showed evaluators in a large-sized district perceived at least three times as many classroom teachers as *Unsatisfactory* or *Developing* than actually rated as such in those categories. When asked to judge what percentage of teachers in their schools perform at an *Unsatisfactory* level, evaluators in their study responded “5% of teachers”, but only 1% of teachers were rated as *Unsatisfactory*. Similarly, when asked what percentage of teachers in their schools perform at a *Developing* level, evaluators responded “15% of teachers”, but only 5% of teachers were rated as *Developing*. Follow-up interviews with evaluators suggested some explanatory factors for the differences between perceived and actual ratings, including evaluators felt an assignment of low ratings may be: (a) unfair if the school does not have the capacity to support a teacher’s improvement; (b) resource-heavy, given additional documentation and observations are needed for low ratings; (c) counterproductive to a teacher’s development, particularly when a teacher accepts feedback and appears motivated to improve; and (d) risky from a personnel standpoint, as the school may lose a good teacher with few candidates to replace the teacher. Whether the findings from the district studied by Kraft and Gilmour generalize to the district in this study is uncertain because SLP evaluators were not interviewed about their assignment of ratings, but communicating weaker performance to a colleague is likely difficult for any peer evaluator. A district administrator of speech-language pathology can support SLP peer evaluators in uncomfortable decisions by insisting a

performance evaluation instrument used for SLPs is valid and reliable (SLPs should feel the ratings are credible, consistent and fair) and ensuring supports for SLPs showing weaker performance.

A third possible reason for the low frequency of *Requiring Attention* and *Developing* performance ratings for SLPs in this study may be due to the preparation of service that goes into planned (announced) evaluations. Announced, compared to unannounced observations, allow educators to show their best work. Data from Ho and Kane (2013) showed performance ratings of announced lessons were .07 points higher in a 4.0-point scale compared to unannounced lessons. Common practice in schools is to include both announced and unannounced observations in performance evaluations (Center on Great Teachers and Leaders, 2014a) and including both types of observations may be a future direction for the district of the 111 SLPs in this study.

High and low performance ratings. SLPs received high ratings for the majority of elements in Domain 2 (Climate of Service), including establishing, maintaining, and explicitly communicating room/space routines and procedures (D2, E2); using effective and constructive behavior management (D2, E3); and building positive relationships with students (D2, E4). More than 75% of SLPs received a rating of *Exemplary* for these elements, much higher than the average percentage of occurrence (41%) for an *Exemplary* rating. This is not the first study to show climate of service for individual or small groups of students may be a relative strength of SLPs. In their study of intervention practices for school-age language impairment, Schmitt et al. (2014) and Biancone et al. (2014) showed an SLP's *Emotional Support* (climate, sensitivity, and regard for student perspectives) and *Proactive Management* (behavior management, productivity, and learning formats) were strong features of intervention sessions.

Although SLP survey participants in this study largely agreed Domain 2 elements represented effective practices for SLPs and the scales for these elements appropriately

placed SLPs on a continuum of performance, providing feedback to SLPs on these high-performing Domain 2 elements (D2, E2; D2, E3; and D2, E4), as the elements are currently written, may not be fruitful for promoting professional growth for SLPs as there is very little room for the SLPs to grow. Future tool developers may wish to modify these elements for further distinction of performance levels, or remove one or more of the elements because the elements minimally discern differences among SLPs.

There were four elements on the SLP performance evaluation instrument that appeared to be areas of development and/or the scales for these elements were written in ways that placed the SLPs in the *Developing* category more often than the average percent of occurrence (9%) for a *Developing* rating. These elements included establishing high expectations for student participation and explicitly setting up the environment so students understand the schedule and purpose of the service session (D2, E1); explicitly communicating speech/language learning objectives (D3, E1); providing evaluations that are appropriate, accurate, and educationally focused (D4, E7); and proposing educational plans that are complete, educationally relevant, and measurable (D4, E8). For these times, ratings of *Developing* were given approximately 30% of the time.

The first two elements (D2, E1 and D3, E1) are similar to each other, in that both involve explicit communication of the intervention session purpose, goals, and activities. This is a feature of evidence-based practice intervention for students with unique needs (Jones & Brownell, 2013) and can be an area growth for professionals serving students with disabilities (Johnson, 2015), particularly when the students have moderate-severe cognitive disabilities (Ruppar et al., 2014). The latter two elements (D4, E7 and D4, E8) are also not unexpected areas of growth for professionals responsible for legal special education compliance, particularly given the frequent changes special education law (Office of Special Education Programs and Rehabilitative Services, 2017) and as a

result, the frequent changes that occur in federal, state, and district special education paperwork (Government Accountability Office, 2016). Although potentially challenging areas of practice, the SLPs in this study agreed the four elements showing lower performance still represented effective practices for SLPs and preliminarily, the implication of these findings suggests it may be important to support SLPs in these areas of growth through professional development and/or the creation of tools and resources. For example, it may be helpful to SLPs to have examples of explicitly communicating learning targets and model special education paperwork forms.

EBP performance ratings. Many SLPs in this study showed proficient performance for the five instrument elements with a specific emphasis on evidence-based practice (EBP); for each of these elements (D1, E1; D1, E6; D3, E5; D4, E10; D4, E11), at least 75% of the SLPs evaluated received a rating of *Proficient* or *Exemplary*. These percentages were higher than expected, given studies have shown EBP implementation occurs 50% of the time in clinical settings (Rangamani, Coppens, Greenwald, & Keintz, 2016). It is possible the indicators of performance for the five EBP elements were less rigorous than the checklists used to measure EBP implementation in the studies reviewed by Rangamani et al. If less rigorous, it may be easier for SLPs to demonstrate "proficiency" on the experimental SLP evaluation instrument, compared to EBP fidelity checklists, thus partly explaining the difference between actual (75%) and predicted (50%) *Proficient* or *Exemplary* ratings. Future tool developers may wish to examine whether the indicators for the five EBP elements on the SLP performance instrument inflated SLP performance.

On the other hand, the ratings for the five EBP elements on the SLP performance evaluation instrument may have reflected actual performance for this single performance evaluation during the 2014-2015 school year. Professional development can prompt instructional changes for SLPs (Mahowald, Lenz, Murray, Pyan, & Rentmeester Disher,

2016) and in 2014-2015, the 111 SLPs in this study had the option of engaging in professional development that is centered around EBP. When asked what professional development courses were offered to SLPs in 2014-2015, the district supplied the following eight course titles: (a) Overview of the EBP Process, (b) Finding Free Online EBP Resources, (c) EBP: Elementary Language-Literacy Intervention, (d) EBP: Adolescent Language-Literacy Intervention, (e) EBP: Speech Sound Disorder Intervention, (f) EBP: The Cognitive, Affective, Linguistic, Motor and Social Assessment for Fluency Disorder, (g) EBP: Core Vocabulary and Aided Language Input for Augmentative and Alternative Communication Users, and (h) EBP: Family-Guided Routines-Based Intervention for Birth-Three Services. While attendance was optional in these courses, attendance logs showed 86 of the 111 SLPs attended at least one of these courses in 2014-2015 and thus, may have used the course content in their 2014-2015 performance evaluations. Whether the EBPs observed in the 2014-2015 performance evaluations were representative of the 111 SLPs' typical practices may be a follow-up research question for the district.

Impact of SLP factors on performance ratings. A linear mixed effects (LME) model was used to assess the potential impact of SLP and session/student variables on evaluator's draft ratings of SLPs. Results showed an SLP's training institution was not a predictor of his/her total performance score. This finding may be due to the standards in place for SLP graduate programs and certification maintenance (Council for Academic Accreditation in Audiology and Speech-Language Pathology of the American Speech-Language-Hearing Association, 2017; Council for Clinical Certification in Audiology and Speech-Language Pathology of the American Speech-Language-Hearing Association, 2013) and/or from the effects of averaging of performance across elements to create a total score (e.g., two SLPs may have the same total score, but each has a different profile of relative strengths and weaknesses across individual elements).

Years of professional experience as an SLP was associated with slightly a lower total performance score, a finding that has also been found for teachers in large-scale performance evaluation research (van de Grift & van der Wal, 2010). In their study of approximately 1,500 classroom teachers in five different countries, van de Grift and van der Wal showed that, on average, a teacher's performance ratings slightly declines after approximately 10 years of service. The impact of this reduced teacher performance on student outcomes was not examined by van de Grift & van der Wal. In the current study, although the impact of SLP professional experience on total performance score was statistically significant, there may be limited practical significance to this finding. With all other SLP and session/student factors held constant, an SLP's total performance score (max of 108 points) was lowered by only 0.20 points for every year experience as an educational SLP.

Impact of Student/Session factors on performance ratings. In terms of session/student variables examined in the LME model, two of the six session/student variables (group status of session and multilingual status of student) were not statistically significant predictors of SLP total performance ratings, providing preliminary evidence the 2014-2015 SLP evaluation instrument and the evaluators were not biased towards or against SLPs serving groups of groups or multilingual students.

Conversely, there were some indications of bias based on the findings for the other four out of six session/student variables. I will begin by reviewing the first three of these variables. With all other SLP and session/student variables controlled, total performance scores were lower by approximately 6-10 points for SLPs serving birth-five ages; SLPs serving students with moderate or severe disabilities; and SLPs serving students with combined language and articulation impairments. There is no theoretical reason why SLPs serving students described by one of these variables would perform lower in a performance evaluation than SLPs serving other student populations, so

preliminarily, these results suggest the experimental SLP performance evaluation instrument should be reviewed to ensure all indicators of performance can be observed for these student/session variations. If, after reviewing the instrument, the tool development team determines the instrument's indicators are observable for these student/session variations, then additional training may be needed for evaluators to recognize these indicator variations for SLPs in future iterations of the performance evaluation instrument.

The last student/session variable statistically associated with lower SLP total performance scores is harder to interpret in terms of potential bias. Results from the LME model showed total performance scores were lower by 9 points for SLPs serving students in speech rooms compared to other locations (classrooms or home settings). It is possible the SLP performance evaluation instrument and/or evaluators are biased against SLPs serving students in speech rooms, but if so, this "bias" may be perceived positively by school districts due to growing evidence of accelerated outcomes for students with disabilities when they have access to typical peers (e.g., Bui, Quirk, Almazan, & Valenti, 2010). For the 2014-2015 SLP performance evaluations, the speech room sessions did not have access to typical peers as part of the sessions.

Rather than bias, it is possible SLPs who served students in pull-out, speech-rooms actually had lower performance compared to their SLP colleagues who served students in more natural settings (homes, community classrooms such as Head Start, or school classrooms). Referring to the elements on the SLP performance evaluation instrument, it is possible students served in speech rooms had fewer functional learning targets (D1, E2) in real-word communication situations (D1, E3) with less connections to what peers are learning in other classroom, home, or community settings (D3, E3).

Research Aim 2

The second research aim included determining the extent to which items on an experimental SLP performance evaluation instrument demonstrated preliminary evidence of construct and face validity.

Construct validity. Preliminary construct validity was estimated through correlational and factor analyses on evaluator draft ratings. The domains on the SLP performance evaluation instrument were developed a priori, based on a classroom teacher rubric used in the district of the 111 SLPs.

Within each domain, many item-item correlations were weak in strength (<.39 correlations), providing preliminary evidence the majority of elements within a domain represented a unique construct. The three highest item-item correlations were for elements in due process paperwork (Domain 4, E7 and E8; .62 correlation), data collection and decision-making (Domain 4, E10 and E11; .59 correlation), and planning coherent, functional service (Domain 1, E2 and E3; .55 correlation). For future development purposes, if seeking a shorter, simpler version of the current SLP performance evaluation instrument, one might consider eliminating an element within each of these pairs, since both elements may be evaluating a similar construct. Eliminating similar items can free up evaluator time so they can track other things during observations (Kane & Staiger, 2012). At the same time, eliminating similar items can conserve SLP time by focusing professional growth efforts on a smaller set of key competencies.

Item-Total correlations demonstrated that some elements were relatively more indicative of an SLP's total score than other elements on the SLP evaluation instrument. Using correlation coefficients of .60 or higher to indicate a moderate-strong relationship, an SLP's planning of intervention that is evidence-based (D1, E1) and functional (D1, E3), and delivery of intervention that is cognitively engaging (D3, E4), monitored for student progress (D3, E5), and designed to promote student independence (D3, E6),

were indicative of an SLP's total score. Additionally, an SLP's quality of evaluation reports (D4, E7), programming of dynamic service delivery (D4, E9), and data-based decision making (D4, E11) were indicative of an SLP's total score. If seeking a more parsimonious version of the current performance evaluation instrument, these elements may not be good candidates to eliminate as they showed a moderate relationship to the overall construct of "SLP quality" as operationalized by the instrument developers.

The exploratory factor analysis (EFA) revealed whether the a priori domains on the SLP performance evaluation instrument made sense according to the actual data collected. Results showed some alignment of instrument elements with a priori domains, but in general, the elements appeared to cluster into an alternative factor structure of four factors that included service planning, service implementation, EBP processes for service, and compliance with legal mandates. Given the SLP performance evaluation instrument under study was modified from a classroom teacher instrument, it may not be surprising the domains of practice for a teacher (Kane & Staiger, 2012) are not the same as the domains of practice for an SLP. Even within classroom instruction, domains of practice for a generalist teacher (e.g., a 4th grade teacher; Kane & Staiger) are not necessarily the same domains of practice for a specialist teacher (e.g., a high school science teacher; Schultz & Pecheone, 2014).

Face validity. Preliminary face validity was estimated by survey feedback from SLPs' evaluation of the SLP performance evaluation instrument and their evaluators. While agreement levels were quite high in both surveys, suggesting the survey participants felt the evaluation instrument represented effective SLP practices and the evaluators provided objective performance feedback, concerns noted in the optional comments of the surveys were aligned with the bias results uncovered in the linear mixed model. That is, not all SLP survey participants felt the indicators applied to all

Birth-22 SLPs and the students they serve, which suggests a review of instrument indicators and/or increased evaluator training.

There are two important notes about the SLP perception surveys. First, the SLPs who responded to the surveys may not have represented all SLPs in the district; lower survey participation rates increase non-response bias. Second, the experimental SLP performance evaluation instrument was not reviewed externally by non-district SLPs, speech-language pathology administrators, or higher-education speech-language pathology faculty. If non-responders and/or larger group of stakeholders had reviewed the experimental SLP performance evaluation instrument, they might have presented higher disagreement ratings and optional comments indicating concerns about the utility, comprehensiveness, or soundness of the SLP performance evaluation instrument.

Research Aim 3

The third research aim included determining the extent to which items on an experimental SLP performance evaluation instrument demonstrated agreement and produced reliable results. Video-taped intervention sessions of 34 SLPs (30% of the sample) were scored by four raters. Generalizability theory was used to identify and measure multiple sources of variance and estimate overall reliability.

SLP variance. In the current study, the percentage of variance due to differences in SLPs was 27% Domain 1, 5% Domain 2, 21% Domain 3, and 17% Domain 4. With the exception of Domain 2, these percentages were similar to the percentages reported for classroom teachers. In special education, 15%-21% of the variance in performance ratings for the evaluation tool, *Recognizing Effective Special Education Teachers* (RESET), was due to differences in teachers (Semmelroth & Johnson, 2014) and in general education, 27%-45% of the variance in performance ratings for the evaluation tool, *Framework for Teaching* (FFT), was due to differences in teachers (Ho & Kane, 2013). Ideally, the variance due to SLPs would be higher, but

given the descriptive results showing a restricted range of performance for several instrument elements, the SLP variance results were not unexpected. As an initial-version instrument, there is room for improvement.

For Domain 2, the variance due to SLPs was minimal (5%), consistent with the descriptive results showing a minimal range of performance for the majority of elements in Domain 2. When there is a minimal range of performance for a sample of participants (a relatively homogenous sample), there is a minimal amount variance that can accounted for by the sample of participants. The variance in Domain 2 performance ratings was primarily accounted for by the items themselves, such that Element 1 was scored differently by the raters than Elements 2, 3, and 4. Both the descriptive and variance decomposition results suggest Domain 2 was limited in its ability to discern differences among the 111 SLPs in this study.

Item effects were also present in the variance decomposition for Domains 1, 3, and 4. However, instead of a large main effect for item, results showed a large interaction effect between SLPs and items. This interaction accounted for 33.8%, 31.9% and 57% of the variance in Domains 1, 3, and 4, respectively, and indicated that some SLPs scored higher on some items within each domain. Stated another way, an SLP had variable performance within each domain; for example, an *Exemplary* rating for one Domain 1 element did not imply an *Exemplary* rating for another Domain 1 element. This SLP x item interaction effects may speak to the variation in clinical skills for an SLP – that is, an SLP may have exemplary skills for some elements, but show developing skills in other areas. Teacher x item interaction effects have been reported for classroom teachers, in which 30% of the variance in performance ratings has been due to this interaction (Praetorius, Pauli, Reusser, Rakoczy, & Klieme, 2014).

Alternatively, the observed SLP x item interaction may be related to the a priori domain structure in the performance evaluation instrument. Correlational and

exploratory factor analyses showed most elements within a domain were only weakly related. If instrument elements were regrouped under an alternative domain structure, it is possible the SLP x item interaction variances would decrease because an *Exemplary* rating for one domain element may imply an *Exemplary* rating for another domain element because they are measuring a similar construct.

Rater variance. The variance decomposition tables showed the percentage of variance due to differences in raters was 0.3% for Domain 1 and 0% for Domains 2-4. The percentages of variance in the performance ratings due to the rater interactions, SLP x Rater and Items x Rater, were also minimal; combined, these interactions accounted for less than 5% of variance in any domain. Minimal rater effects have been reported in other performance evaluation instruments (Hill et al., 2012b; Ho & Kane, 2013; Lawson, 2015; Praetorius et al., 2014; Semmelroth & Johnson, 2014;) and low rater effects are generally a positive feature of a performance evaluation instrument (Hill et al., 2012a).

Residual variance. Although minimal rater effects were noted above, the variance decomposition tables showed sizeable residual error effects. For each domain, the residual error was a combination of the highest interaction effect (SLP x Rater x Items) confounded with random error. For Domains 1, 2, 3, and 4, the percentage of variance due to residual effects was 21%, 13%, 31%, and 16%. This was lower than the percentages (27%-58%) reported for some studies (e.g., Semmelroth & Johnson, 2014) but higher than percentages (2%-22%) reported for other studies (e.g., Praetorius et al., 2014). Like rater effects, low residual effects are generally positive features of a performance evaluation instrument (Hill et al., 2012a).

Optimization. Using .65 as the desired reliability level in either absolute or relative decisions, results of the decision studies showed one evaluator is sufficient for Domains 1 and 4, but two evaluators are needed for Domain 3. Using a more stringent

reliability level of .80 in absolute or relative decisions, at least four evaluators are needed for any domain. Both results are consistent with classroom teacher performance evaluation research, such that one-two raters are needed to achieve an acceptable level of .65 reliability (Ho & Kane, 2013) but three-four raters are needed to achieve a more stringent level of .80 reliability (Ho & Kane, 2013; Kane & Staiger, 2012; Lawson, 2015; Semmelroth & Johnson, 2014).

In sum, examining the results of the study as they pertain to the three research questions broadly suggests the SLP performance evaluation instrument demonstrated acceptable psychometric properties for an initial-version performance evaluation instrument. While there was a restricted range of performance evaluation ratings for SLPs, this was not unlike other studies and the reasons for a restricted range of performance are difficult to untangle (Kraft & Gilmour, 2017). Biases can be expected in observational rubrics (Papay, 2012) and some were evident in this performance evaluation instrument; however, the biases were not so extreme that groups of SLPs appeared to be at a major disadvantage during their performance evaluations. Lastly, an acceptable .65 reliability was reached for two out of the four domains using just one rater, although ideally one rater would be needed to reach .65 reliability for all domains.

Despite some shortcomings in the current version of the performance evaluation instrument, feedback from SLPs indicated they felt the performance evaluation instrument represented effective practices for SLPs and the time spent with their evaluators was valuable. With further tool development, it appears quite possible to improve the psychometric properties of the SLP performance evaluation instrument.

Implications for Practice

This study's investigation of the preliminary validity and reliability evidence for an SLP performance evaluation instrument has implications for school-based SLPs and administrators. Several implications for practice are discussed below.

Purpose of performance evaluation. The selection of educator performance evaluation measures depends on the intended purpose and context of use for a district. Often, there are three main purposes of performance evaluation. One, a performance evaluation system is designed to accurately assess an educator's performance. Two, a performance evaluation system is designed to provide feedback to individual educators, in order to assist them in identifying areas of strength and for areas of growth, provide an actionable pathway to higher levels of performance. Three, by examining patterns of performance across educators, a performance evaluation system is designed to provide feedback to administrator about the professional needs of staff. Early in the design stages of performance evaluation, districts will want to identify the purposes for performance evaluation systems and consider what data should be collected to evaluate these purposes. In the subsequent paragraphs, I assume these three purposes apply to the district of the 111 SLPs in this study and briefly discuss each purpose in light of the study findings.

In terms of the first purpose (accurate evaluations), the absolute and relative G-coefficients from the decisions studies suggest SLPs were rated in comparable and consistent ways for Domains 1 and 4 with one rater (using .65 as the desired reliability level), but two raters were needed to reach that same reliability level for Domain 3. From a resource standpoint, even two evaluators for Domain 3 may not feasible for a district and therefore, additional tool development and evaluator training are needed for Domain 3 in order to reach a higher level of reliability with just one evaluator. For Domain 2, major item development is needed.

In terms of the second purpose (feedback to practitioners), the range of performance for SLPs was restricted in this study, but not all SLPs were rated as *Proficient* or *Exemplary* on all elements, suggesting there is professional room to grow for the SLPs with an actionable roadmap to higher levels of performance. Perception

surveys of the instrument and evaluators demonstrated preliminary indications of credibility and utility – that is, while not all SLPs in the study completed the perception surveys (43% and 65% completion rates), those that did complete the surveys reported the elements on the instrument represented effective SLPs practices and the evaluators provided them objective feedback in a process that was a valuable use of their time.

In terms of the third purpose (feedback to departments), patterns of performance emerged in the data that suggested additional professional development may be needed to support SLPs. More tools and resources may be needed to address the legal compliance of educational evaluations and services plans (D4, E7 and E8) and the explicit communication of student learning targets (D2, E1 and D3, E1).

Importantly, for a speech-language pathology administrator, the LME model results of session/student variables showed the SLP performance evaluation instrument contained some biases, but at this time, not enough severe biases to warrant the creation of separate, parallel versions of a performance evaluation instrument for different student ages, disability areas, and/or contexts of service. Instead of parallel versions, it may be more prudent to review and modify the current performance evaluation instrument to ensure all indicators of performance can be observed for variations in session/student variables and evaluators are adequately trained on these session/student variations.

Limited personnel functions. The purpose of the performance evaluation instrument examined in this study was to assess and develop SLP professional practices, and in turn, promote student learning. Many experimental, initial-version performance evaluation instruments have the same purposes (e.g., Hill et al., 2012a). There is no evidence from this study that suggests this instrument can be used for personnel decisions, such as demotion, promotion, retention, or compensation, as validation tests for those purposes were not included in this study. For example, this

study did not examine whether a cut-off score should be used as the criteria to retain or dismiss an SLP. In fact, the absolute G-coefficients from this study are likely not high enough for any personnel decisions. At best, the absolute G-coefficient for one rater was .67 (Domain 1). Districts may not be comfortable making personnel decisions based on a performance evaluation instrument until the instrument shows at least .80 absolute G-coefficients for all evaluated domains.

Even when performance evaluation cannot be used for personnel decisions, however, it is important not to ignore weak performance during performance evaluation. If poor performance goes unaddressed, performance evaluation becomes limited in its ability to change professional practices and ultimately, student outcomes. Some of the greatest changes in student outcomes have come with the most courageous conversations during performance evaluation (Reinhorn et al., 2016) even when a performance evaluation instrument is still in the initial stages of validation (Johnson, 2015).

Costs and benefits. When the purpose of a performance evaluation instrument is to develop and support staff, and in turn, improve student outcomes, one might consider a performance evaluation system to be a capital investment in staff. The district of the 111 SLPs in this study identified the following costs associated with developing and implementing the 2014-2015 performance evaluation system for SLPs: Initial training for evaluators, booster training for evaluators, stipends for evaluators, professional development for SLPs (in order to learn the learning management system for the district's performance evaluation paperwork), and other administrative costs for evaluators (e.g., supplies, computers, mileage). In return, the district of the 111 SLPs expected the following benefits: Compliance with state law for teacher evaluation, consistency in performance evaluation procedures for all district SLPs, and opportunities for individual SLPs and the speech-language pathology departments to take stock of

professional strengths and needs. Furthermore, by using district SLP peer evaluators, the district promoted leadership and mentorship within the speech-language pathology department (Holdheide et al., 2014). For districts in which SLPs are not considered a teacher under performance evaluation state law and therefore, not required to engage in performance evaluation (Center on Great Teachers and Leaders, 2014), the costs of developing a performance evaluation system may not outweigh the benefits.

Limitations

This study describes validity and reliability evidence for an experimental SLP performance evaluation instrument designed at a local district level. There are at least three limitations in generalizing the findings to other performance evaluation instruments and contexts of use.

"Good" versus "effective" practices. Although perception surveys from SLPs indicated items on the performance evaluation instrument represented effective practices for SLPs, there is no evidence from this study the instrument items are empirically related to student outcomes. Indicators for the highest-level ratings (*Exemplary*) reach toward what is perceived to be top a professional practice, but it is currently unclear if SLPs with *Exemplary* ratings actually produce greater/accelerated student outcomes compared to SLPs with lower performance ratings. Therefore, the descriptor "good" may be a more appropriate term than "effective" to describe the elements and indicators on the SLP performance evaluation instrument. In clinical fields, more advanced "expertise" is believed to be different than core clinical competency (Overholser, 2010) and a performance evaluation instrument should attempt to differentiate between these two practitioner performance levels, but only follow-up validity studies can determine whether distinctions in practitioner performance are associated with higher or lower outcomes for students. Finally, it is possible that different items on the SLP performance evaluation instrument may be associated with

different types of student outcomes. In the general education performance evaluation literature, there is evidence classroom management skills are the strongest predictor of student achievement, but a teacher's caring stance is the strongest predictor of student happiness and sense of self-worth (Ferguson & Danielson, 2014).

Limitations due to evaluators. There was a small number (*N* = 5) of evaluators who evaluated the SLPs in this study. Through training, it possible the group of five developed consensus on scoring that isn't necessarily transferrable to a larger group of evaluators or evaluators outside of the district. Additionally, this study's findings may not be transferrable to districts who utilize content-specific SLPs as evaluators. The SLPs in the study were randomly assigned to one of five evaluators and given this random assignment, there was a possibility of a mismatch between an evaluator's own content expertise and the SLP's evaluated practices. For example, one might imagine feedback to a Birth-Three SLP from a high-school SLP evaluator (Evaluator #3 in this study) has the possibility of being too general and not specific to early intervention practices. Without deep knowledge of Birth-Three evidence-based practices, the high-school SLP evaluator might be hesitant to rate the Birth-Three SLP's skills as *Requires Attention* or *Exemplary*. In classroom teaching, there is evidence that performance ratings remain in the middle of the distribution (e.g., *Proficient* ratings) when evaluators do not have content-specific expertise to judge performance (Rigby et al., 2016). Having a team of evaluators, rather than a single evaluator, increases the probability educators receive differentiated ratings and specific feedback, but an exact content-specific match is not always possible between an SLP and their evaluator.

Similarly, the findings in this study may not be transferrable to districts who use non-SLPs as SLP evaluators. In a recent study by Lawson (2015), the ratings of special education teachers were higher when a school administrator (who did not have formal training or expertise in special education) gave the ratings compared to when a special

education colleague gave the ratings. When asked to evaluate performance this item, "Teacher appears to have a solid understanding of the content", school administrators in scored the item almost twice as high as peer evaluators scored the item. Findings from Lawson (2015) have also been reported in large-scale general education performance evaluation research. Ho and Kane (2013) showed peers evaluators scored general education teachers 0.20 -0.25 points lower on a 4.0 scale, compared to school administrator evaluators.

Inferences about SLP quality. The SLPs in this study were evaluated at a single point in time for a planned (announced) evaluation. Performance ratings from this one evaluation may be an indicator of SLP quality, but not *the* indicator of SLP quality. Evaluations of multiple lessons provide a more reliable assessment of quality than any single evaluation (Hill et al., 2012b; Ho & Kane, 2013; Lawson, 2015; Praetorius et al., 2014; Semmelroth & Johnson, 2014). Moreover, there may be speech-language competencies that affect students that were not measured on the SLP performance evaluation instrument examined in this study.

Future Directions

First and foremost, the data from this study suggests the experimental SLP performance evaluation instrument should continue to undergo the revision and validation process until desired psychometric qualities are established. Continued validity checks are needed until all indicators of performance on the experimental performance evaluation instrument are adequately defined for all student/session variables. Then, even when desired psychometric qualities have been established across items, periodic checks for reliability and bias are needed, as not all performance evaluation instruments remain stable and unbiased overtime (Park et al., 2014). Ongoing statistical monitoring of performance evaluation instruments ensures the

instruments remain appropriate, meaningful, and informative for the staff they aim to support.

Secondly, it may be important to determine if the SLP evaluation instrument examined in this study shows evidence of validity and reliability outside of the district included in this study. Obtaining outside validity and reliability evidence would strengthen the instrument's credibility and utility among SLP professionals. Examining the distribution of performance evaluation ratings in different districts would help determine if ratings on the experimental SLP evaluation instrument are sensitive to contextual factors between districts. If ratings differ across districts, this finding would provide an opportunity to understand why ratings differ. Preliminary data from classroom teacher performance evaluation suggests the availability and type of district-specific resources (e.g., curricula, professional development, instructional supports) may partly explain why teachers in some districts have higher performance ratings than teachers in other districts, even after controlling for teacher and student background characteristics and rating teachers on the same evaluation instrument with the same evaluators (Blazar, Litke, & Barmore, 2016).

Finally, there are two primary outcomes of any performance evaluation system for educators: improve professional practices and in turn, increase student outcomes. Preliminary validity and reliability evidence examined in this study cannot verify either outcome has been met for the 111 SLPs in this study, as the data for this study included archived SLP performance data at a single point in time (i.e., no pre- and post-rating data) and did not include student outcome data. An important future study includes measuring SLPs' growth in performance, following the participation in performance evaluation, and determining whether any that growth in performance is linked to student learning. Increasingly, policymakers need compelling evidence that show all educators

matter in schools and examining the extent to which SLP performance evaluation predicts intervention progress is important future work.

Conclusion

Large-scale educator performance evaluation research has primarily focused on classroom teachers, in part because of the challenges in designing, implementing, and validating performance evaluation instruments for non-classroom educators whose roles and responsibilities are highly variable, specialized based on unique student needs, and grounded in best practices that may be different than effective practices for classroom teachers (Holdheide et al., 2014).

This study examined the validity and reliability evidence for an experimental performance evaluation instrument specifically designed for educational SLPs. The results indicated that some areas of the SLP performance evaluation instrument showed acceptable evidence of preliminary validity and reliability while other areas showed weaker evidence of preliminary validity and reliability. By continuing to examine performance evaluation instruments for SLPs, practitioners, researchers, and policymakers can feel more confident these instruments can be used to improve instructional skills of educational SLPs and in turn, increase outcomes for students. The SLP evaluation instrument examined in this study holds promise as a valid and reliable way to provide constructive feedback to school-based SLPs and support SLPs in their professional growth. The volume of performance evaluation research for general education teachers is quite extensive compared to that for special education and specialized instructional support personnel (Holdheide et al., 2014). The present study contributes to the gap in the literature by conducting performance evaluation research for non-classroom educators.

Table 1

Evaluator (N = 5) demographics in the 2014-2015 SLP performance evaluations

Evaluator	Current Work Setting	Years Experience	Highest Level of Education	ASHA CCC-SLP	Number of SLPs assigned to Evaluator
Evaluator 1	Elementary and Secondary schools	10	Master's	Yes	19
Evaluator 2	Elementary school	24	Master's	Yes	24
Evaluator 3	Secondary schools	18	Master's	Yes	22
Evaluator 4	Early Childhood Special Education	23	Master's	Yes	23
Evaluator 5	Lead SLP	14	Master's	Yes	23

Note. CCC-SLP = Certificate of Clinical Competence in speech-language pathology from the American Speech Language Hearing Association; SLP = speech-language pathologist.

Table 2

Steps of completion for SLPs (N = 111) in their 2014-2015 performance evaluations

Step 1. SLP and Evaluator schedule a 90-minute block of time on the assigned date to include: ● 25 min for Domain 1 examination of artifacts and discussion ● 30 min for Domains 2 and 3 SLP observation with student(s) ● 35 min for Domain 4 examination of artifacts and discussion
Step 2. SLP and Evaluator schedule a 30-minute post-observation conference, one week after the observation.
Step 3. SLP completes preparation questions for Domain 1 at least two days prior to the SLP observation.
Step 4. SLP completes preparation questions for Domain 4 at least two days prior to the SLP observation.
Step 5. SLP and Evaluator meet for the 90-minute block.
Step 6. Evaluator sorts evidences and assigns draft ratings for the 27 elements on the instrument.
Step 7. After the observation but before the post-conference, SLP self-rates performance for the 27 elements on the instrument.
Step 8. SLP completes post-observation reflection questions at least two days prior to the post-observation conference.
Step 9. SLP and Evaluator meet for the post-observation conference to discuss positive aspects of the evaluation, areas of growth, and next steps. SLP and Evaluator finalize ratings based on consensus.
Step 10. SLP signs off on final, consensus ratings.

Note. SLP = speech-language pathologist.

Table 3

Pre-observation questions completed by SLPs (N = 111) in their 2014-2015 performance evaluations

Domain 1
1. Describe if/how Evidence-Based Practice (EBP) influenced your decision of learning targets and/or the structure of today's lesson plan.
2. Describe the general structure of today's lesson plan and when applicable, how today's lesson plan builds on previous and/or future lesson plans.
3. Describe any collaboration with the educational team and parent in selecting today's learning targets.
4. Describe the student's functional (i.e., "real-life", applicable to everyday experiences) communication needs across settings (e.g., speech room, regular education classroom, home, community, etc.).
5. Describe how today's learning targets and lesson plan will relate/carryover to the student's functional communication needs described in question 4.
6. What will the observer see/hear to know you have planned for this session to be culturally relevant and personally meaningful to the student?
7. What relevant resources and technology have you planned for today's session?
8. Where in the session should I be looking for opportunities to assess student learning and how will you use the data you collect today to adapt instruction within today's lesson?
9. Is there anything else you'd like to share about the student and/or your lesson plans for today?
Domain 4
1. How have self-reflection and self-assessment of professional practices influenced your service to students? Please provide a specific example. Additionally, describe any tools and resources you have developed to enhance peer coaching, self-reflection, or self-assessment. Feel free to write "N/A" here:
2. How do you use professional feedback to improve your service to students? Please provide a specific example. Additionally, describe any systems you have in place to gather regular feedback from colleagues, administrators, families, and students.

Table 3 continued

3. How have you helped your colleagues use feedback to improve their services to students? Please provide a specific example.

4. How has professional development influenced your service to students? Please provide a specific example.

Additionally, have you provided any professional learning opportunities for your colleagues? Feel free to write “N/A” here:

5. In what ways have you collaborated with families to improve your service to students? Please provide a specific example.

6. Walk me through the pre-referral process at your school/site.

7. How do you ensure students receive an appropriate continuum of service in the Least Restrictive Environment? Please provide a specific example.

Additionally, describe any tools and resources have you developed for LRE. Feel free to write “N/A” here:

8. Describe your systems (or, show artifacts) for collecting and summarizing data. Please describe any connections between these systems and EBP.

9. How have you used data to make instructional decisions? Please provide a specific example for each of these two areas:

- Assessment
- Intervention

Finally, for each area above, please describe any connections between your instructional decision and EBP.

Note. SLP = speech-language pathologist. For all Domain 1 and 4 questions, the term "student" may include an individual student, a group of students, or family/child and “service” may include assessment and evaluation, intervention and instruction, progress monitoring, and/or consultation and collaboration.

Table 4

Post-observation questions completed by SLPs (N = 111) in their 2014-2015 performance evaluations

1. What do you think went well in your observed student session? What could make the session even stronger? Please provide specific examples and suggestions.
2. How did your assessment strategies provide data of student learning? To what extent did the student achieve his/her learning target(s) in the speech-language session? Provide data that you used to determine student learning.
3. How will you use student performance in the speech-language session to plan future lessons? What are your next steps?

Note. SLP = speech-language pathologist. In post-observation questions, the term "student" may include an individual student, a group of students, or family/child.

Table 5

Summary of assessments for SLPs (N = 111) in their 2014-2015 performance evaluations

Domain	Number of Elements	Method of Assessment	Length of Assessment	Evaluator
Domain 1: Planning of Service	6	Ratings based on portfolio of artifacts and discussion	25 minutes	Trained SLP within the district, randomly assigned to evaluate SLP
Domains 2: Climate of Service and Domain 3: Implementation of Service	4 6	Ratings based on observation of service to student(s)	30-45 minutes	Same evaluator as above
Domain 4: Professional Responsibilities, Due Process Documentation, and Case Management	11	Ratings based on portfolio of artifacts and discussion	35 minutes	Same evaluator as above

Note. SLP = speech-language pathologist.

Table 6

Variance components in a SLP x R x I G-study design

Source	Description
SLP	Variance due to SLPs (object of measurement)
R	Variance due to raters
I	Variance due to instrument items
SLP x R	Some raters score some SLPs higher or lower than others
SLP x I	Some SLPs score higher or lower on some items than on other items
R x I	Some raters score some items higher or lower than other items
SLP x R x I, *e*	Highest order interaction, confounded with residual error

Note. SLP = speech-language pathologist.

Table 7

Frequency count and direction of change from evaluator draft to final consensus ratings of speech-language pathologists (N = 111) in 2014-2015 performance evaluations

Element	RA to DEV	DEV to PRO	PRO to EXE	EXE to PRO	PRO to DEV	DEV to RA	*n* Changes
D1, E1		2	2				4
D1, E2			2				2
D1, E3		2	2		1		5
D1, E4			1				1
D1, E5			2	1			3
D1, E6			2				2
D2, E1		1	1				2
D2, E2			1				1
D2, E3			1				1
D2, E4							0
D3, E1		1					1
D3, E2		1	2				3
D3, E3				1			1
D3, E4			1	1			2
D3, E5		2		1			3
D3, E6		1	2	1			4
D4, E1							0
D4, E2		1	1				2
D4, E3							0
D4, E4							0
D4, E5		1	2				3
D4, E6	1	1	1		1		4
D4, E7		1					1
D4, E8	1	2					3
D4, E9		2	2				4
D4, E10							0
D4, E11		1			1		2

Note. RA = Requires Attention; DEV=Developing; PRO=Proficient; EXE=Exemplary; D1,E1= Uses knowledge of Evidence-Based Practice to select learning targets and plan service; D1,E2 = Designs coherent, sequential speech/language sessions that are aligned with student needs; D1,E3 = Designs speech/language sessions that place learning targets in "real-life", functional contexts; D1,E4 = Designs speech/language sessions that are culturally relevant and personally meaningful; D1,E5 = Plans for relevant resources; D1,E6 = Plans for assessment strategies to monitor student progress; D2,E1 = Establishes high expectations for student participation and explicitly sets up the environment so students understand the schedule and purpose of the service session; D2,E2 = Establishes, maintains, and explicitly communicates room/space routines and procedures; D2,E3 = Uses effective and constructive behavior management;

Table 7 continued

D2,E4 = Builds positive relationships (i.e., rapport) with students; D3,E1 = Explicitly communicates speech/language learning objectives; D3,E2 = Provides learning activities and uses instructional strategies that are engaging and motivating; D3,E3 = Provides opportunities to practice skills and demonstrate learning at a rigorous pace while ensuring student success; D3,E4 = Uses techniques to promote cognitive engagement with learning targets; D3,E5 = Uses assessment strategies to monitor student progress; D3,E6 = Gives explicit and timely feedback to improve student learning; D4,E1 = Uses self-reflection and self-assessment to improve service to students; D4,E2 = Uses feedback to improve service to students; D4,E3 = Participates in relevant professional development; D4,E4 = Collaborates with educational team; D4,E5 = Collaborates with families; D4,E6 = Applies knowledge of the pre-referral process in the prevention and identification of disabilities; D4,E7 = Provides evaluations that are appropriate, accurate, and educationally-focused; D4,E8 = Proposes educational plans that are complete, educationally relevant, and measurable; D4,E9 = Provides dynamic service delivery in the least restrictive environment; D4,E10 = Establishes data collection systems for all provided services; D4,E11 = Makes data-based decisions for all provided services

Table 8

Percentage of occurrence for student and session variables self-reported by speech-language pathologists (N = 111) in their 2014-2015 observations

Variable	%
Student Grade	
Birth-5	26.13
Primary School	51.35
Secondary School	22.52
Student Linguistic Status	
Monolingual English	59.46
Bilingual	40.54
Disorder Addressed in Session	
Language	50.45
Articulation	25.23
Language and Articulation	21.62
Fluency	2.70
Perceived Severity of Disorder	
Mild	9.91
Moderate	44.14
Severe	45.95
Session Location	
Home	6.31
Speech Room	75.68
Classroom	18.02
Session Grouping	
Individual Service	74.77
Group Service	25.23

Note. Bilingual students included Spanish-English (48.88%), Somali-English (24.44%), Hmong-English (11.11%), and Other Bilingual (15.57%) students.

Table 9

Linear mixed effects (LME) model coefficients for fixed effects in evaluator draft ratings of speech-language pathologists (N = 111) in 2014-2015 performance evaluations

Variable	Estimate	Standard Error	Degrees Freedom	*t* value	*p* value
(Intercept)	93.69	11.85	94.70	7.91	*p*<.001 ***
Every year of SLP professional experience	-0.18	0.08	95.56	-2.29	.02 *
SLP training institution-Research	-1.72	1.71	95.13	-0.99	.32
Student Age-Primary	10.02	3.97	95.04	2.52	.01 **
Student Age-Secondary	10.70	4.11	95.19	2.60	.01 **
Student-Multilingual	-0.41	1.54	95.00	-0.26	.79
Student Disorder-Language	-2.21	1.94	94.56	-1.13	.26
Student Disorder-Combined Language & Articulation	-6.04	2.29	92.80	-2.63	.009 **
Student Disorder-Fluency	6.36	4.99	95.56	1.27	.21
Student Disorder Severity-Moderate	-8.15	4.05	94.25	-2.08	.04 *
Student Disorder Severity-Severe	-8.23	3.73	94.82	-2.20	.03 *
Session Location-Home	0.81	5.99	93.25	0.14	.89
Session Location-Speech Room	-9.32	3.11	95.81	-2.99	.003 **
Session-Individual	-1.45	1.83	93.24	-0.79	.43
Session Length	0.31	0.33	94.78	0.95	.35

Note. SLP = speech-language pathologist.

*p<.05 **p<.01 ***p<.001

Table 10

Domain 1: Bivariate Spearman coefficients for item-item and item-total correlations (N = 111) in evaluator draft ratings of speech-language pathologists (N = 111) in 2014-2015 performance evaluations

Element	1	2	3	4	5	6	7
1. D1,E1	1.0						
2. D1,E2	.33**	1.0					
3. D1,E3	.37**	.55**	1.0				
4. D1,E4	.26**	.29**	.42**	1.0			
5. D1,E5	.39**	.32**	.27**	.12	1.0		
6. D1,E6	.37**	.26**	.21*	.22*	.27**	1.0	
7. Total Score	.66**	.56**	.62**	.41**	.57**	.44**	1.0

Note. D1,E1= Uses knowledge of Evidence-Based Practice to select learning targets and plan service; D1,E2 = Designs coherent, sequential speech/language sessions that are aligned with student needs; D1,E3 = Designs speech/language sessions that place learning targets in "real-life", functional contexts; D1,E4 = Designs speech/language sessions that are culturally relevant and personally meaningful; D1,E5 = Plans for relevant resources; D1,E6 = Plans for assessment strategies to monitor student progress.

*p<.05 **p<.01 ***p<.001

Table 11

Domain 2: Bivariate Spearman coefficients for item-item and item-total correlations (N = 111) in evaluator draft ratings of speech-language pathologists (N = 111) in 2014-2015 performance evaluations

Element	1	2	3	4	5
1. D2,E1	1.0				
2. D2,E2	.27**	1.0			
3. D2,E3	.26**	.48**	1.0		
4. D2,E4	.08	.37**	.42**	1.0	
5. Total Score	.52**	.47**	.32**	.21*	1.0

Note. D2,E1 = Establishes high expectations for student participation and explicitly sets up the environment so students understand the schedule and purpose of the service session; D2,E2 = Establishes, maintains, and explicitly communicates room/space routines and procedures; D2,E3 = Uses effective and constructive behavior management; D2,E4 = Builds positive relationships (i.e., rapport) with students.

*p<.05 **p<.01 ***p<.001

Table 12

Domain 3: Bivariate Spearman coefficients for item-item and item-total correlations (N = 111) in evaluator draft ratings of speech-language pathologists (N = 111) in 2014-2015 performance evaluations

Element	1	2	3	4	5	6	7
1. D3,E1	1.0						
3. D3,E2	.20	1.0					
3. D3,E3	.13	.32**	1.0				
4. D3,E4	.35**	.42**	.31**	1.0			
5. D3,E5	.36**	.26**	.50**	.35**	1.0		
6. D3,E6	.17*	.39**	.30**	.48**	.34**	1.0	
7. Total Score	.51* *	.55**	.55**	.67**	.63**	.60**	1.0

Note. D3,E1 = Explicitly communicates speech/language learning objectives; D3,E2 = Provides learning activities and uses instructional strategies that are engaging and motivating; D3,E3 = Provides opportunities to practice skills and demonstrate learning at a rigorous pace while ensuring student success; D3,E4 = Uses techniques to promote cognitive engagement with learning targets; D3,E5 = Uses assessment strategies to monitor student progress; D3,E6 = Gives explicit and timely feedback to improve student learning.

*p<.05 **p<.01 ***p<.001

Table 13

Domain 4: Bivariate Spearman coefficients for item-item and item-total correlations (N = 111) in evaluator draft ratings of speech-language pathologists (N = 111) in 2014-2015 performance evaluations

Element	1	2	3	4	5	6	7	8	9	10	11	12
1. D4,E1	1.0											
2. D4,E2	.22*	1.0										
3. D4,E3	.16	.35**	1.0									
4. D4,E4	.12	.13	.09	1.0								
5. D4,E5	.09	.18	.22*	.23*	1.0							
6. D4,E6	.23*	.06	-.02	.12	.04	1.0						
7. D4,E7	.20*	.06	.15	.21*	.30**	.35**	1.0					
8. D4,E8	.19*	.06	.12	.03	.40**	.28**	.62**	1.0				
9. D4,E9	.23*	.28**	.32**	.18	.46**	.14	.40**	.39**	1.0			
10. D4,E10	.26**	.27**	.14	.08	.21*	.03	.22*	.17	.35**	1.0		
11. D4,E11	.22*	.23*	.18	.11	.36**	.17*	.37**	.40**	.46**	.59**	1.0	
12. Total Score	.39**	.34**	.29**	.30**	.57**	.34**	.60**	.56**	.62**	.43**	.61**	1.0

Note. D4,E1 = Uses self-reflection and self-assessment to improve service to students; D4,E2 = Uses feedback to improve service to students; D4,E3 = Participates in relevant professional development; D4,E4 = Collaborates with educational team; D4,E5 = Collaborates with families; D4,E6 = Applies knowledge of the pre-referral process in the prevention and identification of disabilities; D4,E7 = Provides evaluations that are appropriate, accurate, and educationally-focused; D4,E8 = Proposes educational plans that are complete, educationally relevant, and measurable; D4,E9 = Provides dynamic service delivery in the least restrictive environment; D4,E10 = Establishes data collection systems for all provided services; D4,E11 = Makes data-based decisions for all provided services.

*p<.05 **p<.01 ***p<.001

Table 14

Eigenvalues for the full-factor model (27 elements as components) in the exploratory factor analysis of evaluator draft ratings of speech-language pathologists (N = 111) in 2014-2015 performance evaluations

Eigenvalues			
Component	Value	% Variance	% Cumulative Variance
1	7.661	28.374	28.374
2	2.155	7.982	36.356
3	1.675	6.205	42.561
4	1.596	5.911	48.472
5	1.318	4.883	53.355
6	1.162	4.302	57.657
7	1.114	4.125	61.783
8	1.025	3.797	65.580
9	0.937	3.471	69.050
10	0.844	3.125	72.176
11	0.764	2.830	75.005
12	0.729	2.702	77.707
13	0.694	2.571	80.278
14	0.630	2.332	82.611
15	0.588	2.177	84.788
16	0.546	2.022	86.810
17	0.474	1.755	88.565
18	0.455	1.686	90.251
19	0.436	1.613	91.864
20	0.379	1.403	93.267
21	0.366	1.354	94.621
22	0.311	1.153	95.774
23	0.291	1.079	96.853
24	0.258	0.956	97.809
25	0.237	0.876	98.685
26	0.191	0.707	99.393
27	0.164	0.607	100.000

Table 15

Factor loadings for the four-factor model in the exploratory factor analysis of evaluator draft ratings of speech-language pathologists (N = 111) in 2014-2015 performance evaluations

Rotated Factor Matrix				
	Factor			
Element	1	2	3	4
D3, E2	**.629**	-.005	**.328**	.138
D2, E2	**.621**	.121	-.053	.297
D2, E3	**.587**	.073	-.004	.028
D3, E3	**.484**	.179	.186	.260
D3, E6	**.483**	.161	.195	**.368**
D3, E4	**.441**	**.317**	**.389**	.238
D1, E5	**.403**	.291	.116	**.310**
D2, E4	**.395**	.040	.076	.107
D4, E11	.009	**.568**	.139	**.560**
D1, E1	.066	**.534**	**.462**	.294
D4, E10	-.074	**.524**	.082	.296
D2, E1	**.483**	**.515**	.000	-.079
D3, E1	.285	**.506**	.068	.061
D3, E5	**.487**	**.501**	.027	.219
D1, E6	.196	**.463**	.217	.016
D4, E2	-.111	**.420**	**.379**	-.014
D4, E4	.143	.264	.090	.087
D4, E8	.201	.054	.150	**.723**
D4, E7	.235	.110	.134	**.692**
D4, E6	.162	.087	.052	**.312**
D1, E4	.174	.023	**.613**	.032
D4, E3	-.125	.178	**.548**	.108
D1, E3	**.390**	.094	**.545**	.109
D4, E9	.178	.218	**.470**	**.430**
D1, E2	**.410**	.153	**.429**	.156
D4, E5	.299	.180	**.386**	.284
D4, E1	.069	.242	.261	.171

Note. Factor loadings > .30 are boldface. D1,E1= Uses knowledge of Evidence-Based Practice to select learning targets and plan service; D1,E2 = Designs coherent, sequential speech/language sessions that are aligned with student needs; D1,E3 = Designs speech/language sessions that place learning targets in "real-life", functional contexts; D1,E4 = Designs speech/language sessions that are culturally relevant and personally meaningful; D1,E5 = Plans for relevant resources;

Table 15 continued

D1,E6 = Plans for assessment strategies to monitor student progress; D2,E1 = Establishes high expectations for student participation and explicitly sets up the environment so students understand the schedule and purpose of the service session; D2,E2 = Establishes, maintains, and explicitly communicates room/space routines and procedures; D2,E3 = Uses effective and constructive behavior management; D2,E4 = Builds positive relationships (i.e., rapport) with students; D3,E1 = Explicitly communicates speech/language learning objectives; D3,E2 = Provides learning activities and uses instructional strategies that are engaging and motivating; D3,E3 = Provides opportunities to practice skills and demonstrate learning at a rigorous pace while ensuring student success; D3,E4 = Uses techniques to promote cognitive engagement with learning targets; D3,E5 = Uses assessment strategies to monitor student progress; D3,E6 = Gives explicit and timely feedback to improve student learning; D4,E1 = Uses self-reflection and self-assessment to improve service to students; D4,E2 = Uses feedback to improve service to students; D4,E3 = Participates in relevant professional development; D4,E4 = Collaborates with educational team; D4,E5 = Collaborates with families; D4,E6 = Applies knowledge of the pre-referral process in the prevention and identification of disabilities; D4,E7 = Provides evaluations that are appropriate, accurate, and educationally-focused; D4,E8 = Proposes educational plans that are complete, educationally relevant, and measurable; D4,E9 = Provides dynamic service delivery in the least restrictive environment; D4,E10 = Establishes data collection systems for all provided services; D4,E11 = Makes data-based decisions for all provided services.

Table 16

Percentage of occurrence for agreement ratings by speech-language pathologists (n = 47) in their evaluation of 2014-2015 instrument elements

Item	% Strongly Disagree	% Disagree	% Agree	% Strongly Agree
D1, E1	0.00	0.00	44.68	55.32
D1, E2	0.00	0.00	19.15	80.85
D1, E3	0.00	0.00	48.94	51.06
D1, E4	0.00	2.13	44.68	53.19
D1, E5	2.13	4.26	53.19	40.43
D1, E6	2.13	2.13	40.43	55.32
D2, E1	0.00	0.00	34.04	65.96
D2, E2	2.13	4.26	48.94	44.68
D2, E3	2.13	0.00	48.94	48.94
D2, E4	2.13	8.51	29.79	68.09
D3, E1	0.00	0.00	36.17	53.19
D3, E2	0.00	0.00	29.79	70.21
D3, E3	0.00	2.13	46.81	53.19
D3, E4	0.00	2.13	61.70	36.17
D3, E5	0.00	0.00	51.06	48.94
D3, E6	0.00	2.13	31.91	65.96
D4, E1	2.13	2.13	38.30	57.45
D4, E2	2.13	4.26	46.81	46.81
D4, E3	2.13	4.26	34.04	59.57
D4, E4	0.00	2.13	34.04	63.83
D4, E5	0.00	0.00	51.06	48.94
D4, E6	2.13	4.26	61.70	31.91
D4, E7	0.00	0.00	29.79	70.21
D4, E8	0.00	2.13	27.66	70.21
D4, E9	0.00	4.26	57.45	38.30
D4, E10	0.00	2.13	48.94	48.94
D4, E11	0.00	2.13	36.17	61.70
Average% (SD)	0.71 (1.02)	2.05 (2.08)	42.08 (10.83)	55.16 (11.85)

Note. D1,E1= Uses knowledge of Evidence-Based Practice to select learning targets and plan service; D1,E2 = Designs coherent, sequential speech/language sessions that are aligned with student needs; D1,E3 = Designs speech/language sessions that place learning targets in "real-life", functional contexts; D1,E4 = Designs speech/language sessions that are culturally relevant and personally meaningful; D1,E5 = Plans for relevant resources; D1,E6 = Plans for assessment strategies to monitor student progress; D2,E1 = Establishes high expectations for student participation and explicitly sets up the environment so students understand the schedule and purpose of the service session; D2,E2 = Establishes, maintains, and explicitly communicates room/space routines and procedures; D2,E3 = Uses effective and constructive behavior management; D2,E4 = Builds positive relationships (i.e., rapport) with students; D3,E1 = Explicitly communicates speech/language learning objectives; D3,E2 = Provides learning activities and uses instructional strategies that are engaging and motivating; D3,E3 = Provides opportunities to practice skills and demonstrate learning at a rigorous pace while ensuring student success; D3,E4 = Uses techniques to promote cognitive engagement with learning targets; D3,E5 = Uses assessment strategies to monitor student progress; D3,E6 = Gives explicit and timely feedback to improve student learning; D4,E1 = Uses self-reflection and self-assessment to improve service to students; D4,E2 = Uses feedback to improve service to students; D4,E3 = Participates in relevant professional development;

Table 16 continued

D4,E4 = Collaborates with educational team; D4,E5 = Collaborates with families; D4,E6 = Applies knowledge of the pre-referral process in the prevention and identification of disabilities; D4,E7 = Provides evaluations that are appropriate, accurate, and educationally-focused; D4,E8 = Proposes educational plans that are complete, educationally relevant, and measurable; D4,E9 = Provides dynamic service delivery in the least restrictive environment; D4,E10 = Establishes data collection systems for all provided services; D4,E11 = Makes data-based decisions for all provided services.

Table 17

Percentage of occurrence for agreement ratings by speech-language pathologists (n = 47) in their evaluation of 2014-2015 instrument indicators

Item	% Strongly Disagree	% Disagree	% Agree	% Strongly Agree
D1, E1 indicators	2.13	8.51	46.81	42.55
D1, E2 indicators	0.00	0.00	38.30	61.70
D1, E3 indicators	0.00	14.89	51.06	34.04
D1, E4 indicators	2.13	8.51	53.19	36.17
D1, E5 indicators	2.13	6.38	57.45	34.04
D1, E6 indicators	4.26	14.89	51.06	29.79
D2, E1 indicators	0.00	10.64	40.43	48.94
D2, E2 indicators	2.13	6.38	55.32	36.17
D2, E3 indicators	2.13	2.13	53.19	42.55
D2, E4 indicators	2.13	4.26	34.04	59.57
D3, E1 indicators	2.13	12.77	38.30	46.81
D3, E2 indicators	0.00	2.13	44.68	53.19
D3, E3 indicators	0.00	4.26	48.94	46.81
D3, E4 indicators	0.00	4.26	63.83	31.91
D3, E5 indicators	2.13	0.00	59.57	38.30
D3, E6 indicators	0.00	0.00	46.81	53.19
D4, E1 indicators	2.13	6.38	53.19	38.30
D4, E2 indicators	2.13	6.38	55.32	36.17
D4, E3 indicators	2.13	6.38	48.94	42.55
D4, E4 indicators	2.13	2.13	40.43	55.32
D4, E5 indicators	2.13	2.13	53.19	42.55
D4, E6 indicators	4.26	4.26	65.96	25.53
D4, E7 indicators	0.00	0.00	40.43	59.57
D4, E8 indicators	0.00	2.13	36.17	61.70
D4, E9 indicators	0.00	4.26	63.83	31.91
D4, E10 indicators	0.00	4.26	51.06	44.68
D4, E11 indicators	0.00	2.13	46.81	51.06
Average% (SD)	1.34 (1.34)	5.20 (4.27)	49.57 (8.63)	43.89 (10.36)

Note. D1,E1= Uses knowledge of Evidence-Based Practice to select learning targets and plan service; D1,E2 = Designs coherent, sequential speech/language sessions that are aligned with student needs; D1,E3 = Designs speech/language sessions that place learning targets in "real-life", functional contexts; D1,E4 = Designs speech/language sessions that are culturally relevant and personally meaningful; D1,E5 = Plans for relevant resources; D1,E6 = Plans for assessment strategies to monitor student progress; D2,E1 = Establishes high expectations for student participation and explicitly sets up the environment so students understand the schedule and purpose of the service session; D2,E2 = Establishes, maintains, and explicitly communicates room/space routines and procedures; D2,E3 = Uses effective and constructive behavior management; D2,E4 = Builds positive relationships (i.e., rapport) with students; D3,E1 = Explicitly communicates speech/language learning objectives; D3,E2 = Provides learning activities and uses instructional strategies that are engaging and motivating; D3,E3 = Provides opportunities to practice skills and demonstrate learning at a rigorous pace while ensuring student success; D3,E4 = Uses techniques to promote cognitive engagement with learning targets; D3,E5 = Uses assessment strategies to monitor student progress; D3,E6 = Gives explicit and timely feedback to improve student learning;

Table 17 continued

D4,E1 = Uses self-reflection and self-assessment to improve service to students; D4,E2 = Uses feedback to improve service to students; D4,E3 = Participates in relevant professional development; D4,E4 = Collaborates with educational team; D4,E5 = Collaborates with families; D4,E6 = Applies knowledge of the pre-referral process in the prevention and identification of disabilities; D4,E7 = Provides evaluations that are appropriate, accurate, and educationally-focused; D4,E8 = Proposes educational plans that are complete, educationally relevant, and measurable; D4,E9 = Provides dynamic service delivery in the least restrictive environment; D4,E10 = Establishes data collection systems for all provided services; D4,E11 = Makes data-based decisions for all provided services.

Table 18

Percentage of occurrence for agreement ratings by speech-language pathologists (n = 73) in their evaluation of their 2014-2015 evaluators (N = 5)

Survey Question	% Strongly Disagree	% Disagree	% Agree	% Strongly Agree
My observer provided me with specific feedback.	0.00	2.00	20.00	78.00
My observer provided clear evidence of my performance aligned to the rubric.	1.37	1.37	23.29	74.00
The time I spent with this observer was a valuable use of my time.	1.37	2.74	30.14	65.75
I was able to have a reflective two-way dialogue with this observer.	1.37	3.00	15.00	81.00
I felt respected as a professional with this observer during the evaluation process.	1.37	1.37	9.59	88.00
Average% (SD)	1.10 (.61)	2.10 (.76)	19.60 (7.84)	77.35 (8.26)

Table 19

Percentage of occurrence for student and session variables self-reported by speech-language pathologists in the videotaped observations used for 2014-2015 interrater reliability (n = 34)

Variable	%
Student Grade	
Birth-5	14.71
Primary School	55.88
Secondary School	29.41
Student Linguistic Status	
Monolingual English	58.82
Bilingual	41.18
Disorder Addressed in Session	
Language	50.00
Articulation	17.65
Language and Articulation	26.47
Fluency	5.88
Perceived Severity of Disorder	
Mild	2.94
Moderate	55.88
Severe	41.18
Session Location	
Home	5.88
Speech Room	88.24
Classroom	5.88
Session Grouping	
Individual Service	76.47
Group Service	23.53

Note. Bilingual students included Spanish-English (35.72%), Somali-English (28.57%), Hmong-English (21.43%), and Other Bilingual (14.28%) students.

Table 20

Variance decomposition in the videotaped performance evaluations used for interrater reliability (n = 34)

	Domains			
Sources of Variation	1	2	3	4
SLP (s)	27.4	5.0	21.1	17.5
Rater (r)	0.0	0.0	0.3	0.0
Item (i)	16.6	63.7	10.6	8.9
SLP x Rater (s x r)	1.0	1.4	4.9	0.7
SLP x Item (s x i)	33.8	16.7	31.9	57.0
Rater x Item (r x i)	0.6	0.0	0.0	0.1
SLP x Rater x Item (s x r x i), Residual Error (*e*)	20.6	13.1	31.3	15.9
Total Variance	100	100	100	100

Note. Cells represent the percentage of variance. SLP = speech-language pathologist; Domain 1 = Planning of Service; Domain 2 = Climate of Service; Domain 3 = Implementation of Service; Domain 4 = Professional Responsibilities, Due Process Documentation, and Case Management.

Table 21

Decision-study: Absolute and relative G-coefficients with standard error of measurement (SEM) for comparing number of raters in 2014-2015 SLP performance evaluations

Condition	Absolute G-Coefficient	Absolute SEM	Relative G-Coefficient	Relative SEM
Domain 1				
One rater	.68	.20	.73	.18
Two raters	.72	.19	.78	.16
Three raters	.73	.18	.79	.15
Four raters	.74	.18	.80	.15
Domain 2				
One rater	.17	.40	.36	.24
Two raters	.18	.38	.43	.21
Three raters	.19	.38	.47	.19
Four raters	.19	.37	.48	.19
Domain 3				
One rater	.55	.26	.58	.25
Two raters	.63	.22	.67	.20
Three raters	.67	.21	.71	.19
Four raters	.69	.20	.73	.18
Domain 4				
One rater	.67	.18	.69	.17
Two raters	.69	.17	.72	.16
Three raters	.70	.16	.73	.15
Four raters	.71	.16	.74	.15

Note. SLP = speech-language pathologist; Domain 1 = Planning of Service; Domain 2 = Climate of Service; Domain 3 = Implementation of Service; Domain 4 = Professional Responsibilities, Due Process Documentation, and Case Management.

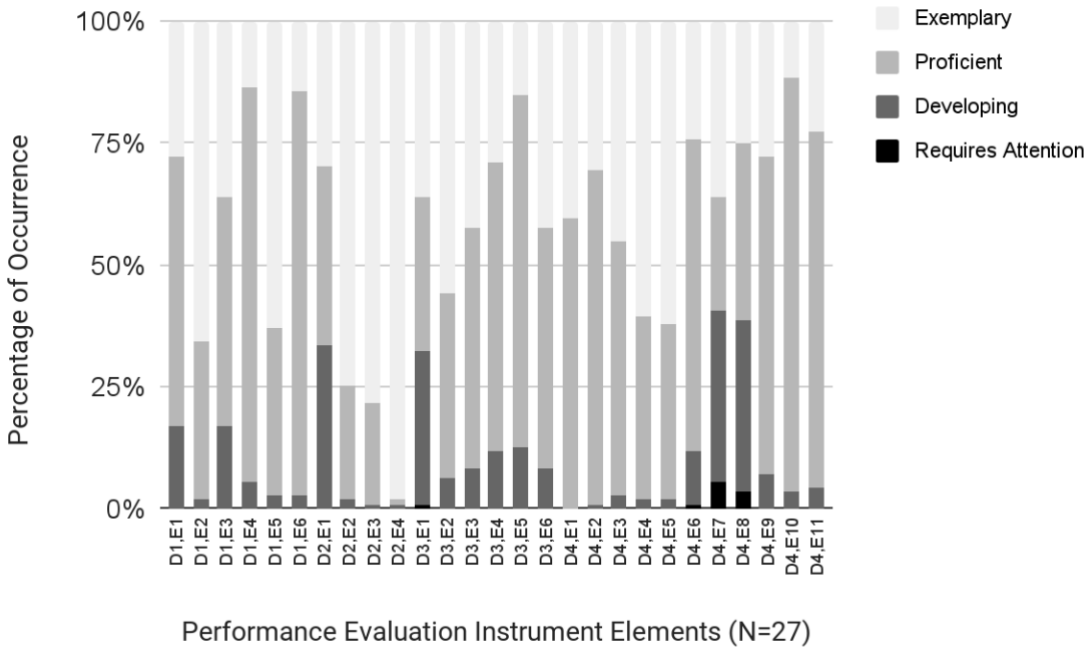

Figure 1. Distribution of evaluator draft ratings of speech-language pathologists (*N* = 111) in 2014-2015 performance evaluations

Note. D1,E1= Uses knowledge of Evidence-Based Practice to select learning targets and plan service; D1,E2 = Designs coherent, sequential speech/language sessions that are aligned with student needs; D1,E3 = Designs speech/language sessions that place learning targets in "real-life", functional contexts; D1,E4 = Designs speech/language sessions that are culturally relevant and personally meaningful; D1,E5 = Plans for relevant resources; D1,E6 = Plans for assessment strategies to monitor student progress; D2,E1 = Establishes high expectations for student participation and explicitly sets up the environment so students understand the schedule and purpose of the service session; D2,E2 = Establishes, maintains, and explicitly communicates room/space routines and procedures; D2,E3 = Uses effective and constructive behavior management; D2,E4 = Builds positive relationships (i.e., rapport) with students; D3,E1 = Explicitly communicates speech/language learning objectives; D3,E2 = Provides learning activities and uses instructional strategies that are engaging and motivating; D3,E3 = Provides opportunities to practice skills and demonstrate learning at a rigorous pace while ensuring student success; D3,E4 = Uses techniques to promote cognitive engagement with learning targets; D3,E5 = Uses assessment strategies to monitor student progress; D3,E6 = Gives explicit and timely feedback to improve student learning; D4,E1 = Uses self-reflection and self-assessment to improve service to students; D4,E2 = Uses feedback to improve service to students; D4,E3 = Participates in relevant professional development; D4,E4 = Collaborates with educational team; D4,E5 = Collaborates with families; D4,E6 = Applies knowledge of the pre-referral process in the prevention and identification of disabilities; D4,E7 = Provides evaluations that are appropriate, accurate, and educationally-focused; D4,E8 = Proposes educational plans that are complete, educationally relevant, and measurable; D4,E9 = Provides dynamic service delivery in the least restrictive environment; D4,E10 = Establishes data collection systems for all provided services; D4,E11 = Makes data-based decisions for all provided services.

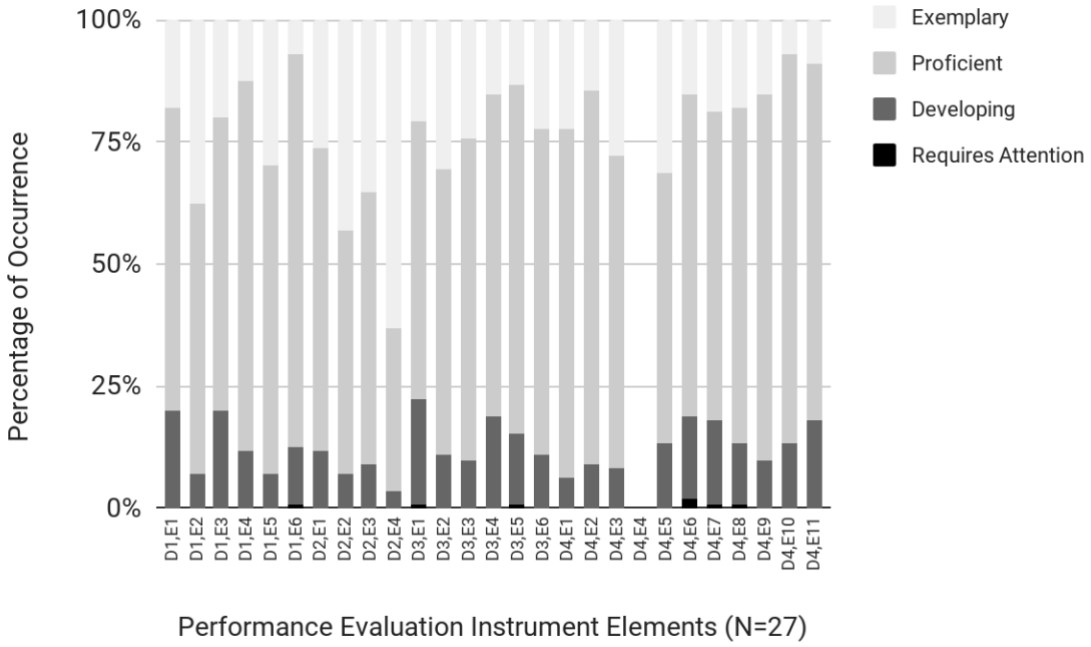

Figure 2. Distribution of self-ratings of speech-language pathologists (*N* = 111) in 2014-2015 performance evaluations.

Note. D1,E1= Uses knowledge of Evidence-Based Practice to select learning targets and plan service; D1,E2 = Designs coherent, sequential speech/language sessions that are aligned with student needs; D1,E3 = Designs speech/language sessions that place learning targets in "real-life", functional contexts; D1,E4 = Designs speech/language sessions that are culturally relevant and personally meaningful; D1,E5 = Plans for relevant resources; D1,E6 = Plans for assessment strategies to monitor student progress; D2,E1 = Establishes high expectations for student participation and explicitly sets up the environment so students understand the schedule and purpose of the service session; D2,E2 = Establishes, maintains, and explicitly communicates room/space routines and procedures; D2,E3 = Uses effective and constructive behavior management; D2,E4 = Builds positive relationships (i.e., rapport) with students; D3,E1 = Explicitly communicates speech/language learning objectives; D3,E2 = Provides learning activities and uses instructional strategies that are engaging and motivating; D3,E3 = Provides opportunities to practice skills and demonstrate learning at a rigorous pace while ensuring student success; D3,E4 = Uses techniques to promote cognitive engagement with learning targets; D3,E5 = Uses assessment strategies to monitor student progress; D3,E6 = Gives explicit and timely feedback to improve student learning; D4,E1 = Uses self-reflection and self-assessment to improve service to students; D4,E2 = Uses feedback to improve service to students; D4,E3 = Participates in relevant professional development; D4,E5 = Collaborates with families; D4,E6 = Applies knowledge of the pre-referral process in the prevention and identification of disabilities; D4,E7 = Provides evaluations that are appropriate, accurate, and educationally-focused; D4,E8 = Proposes educational plans that are complete, educationally relevant, and measurable; D4,E9 = Provides dynamic service delivery in the least restrictive environment; D4,E10 = Establishes data collection systems for all provided services; D4,E11 = Makes data-based decisions for all provided services.

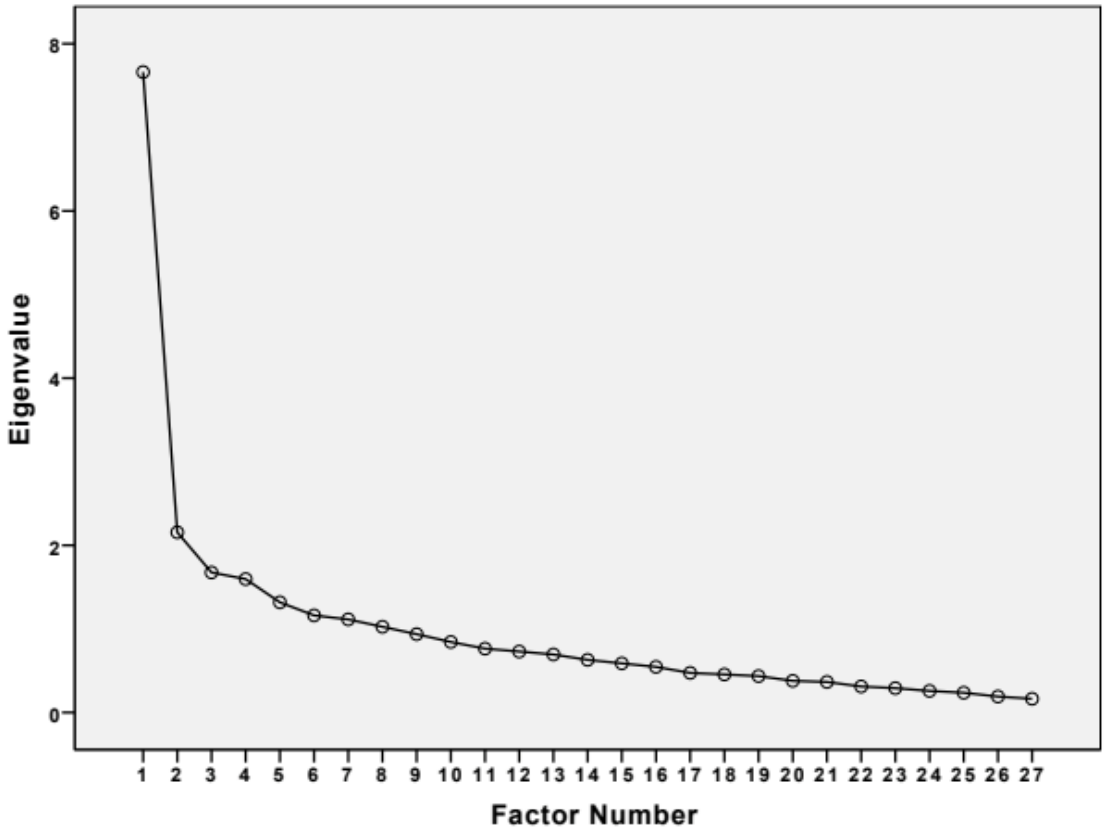

Figure 3. Scree plot of eigenvalues for the full-factor model (27 elements as components) in the exploratory factor analysis (*N* = 111).

REFERENCES

Allen, J. P., Hafen, C. A., Gregory, A. C., Mikami, A. Y., & Pianta, R. (2015). Enhancing secondary school instruction and student achievement: Replication and extension of the My Teaching Partner–Secondary intervention. *Journal of Research on Educational Effectiveness, 8*(4), 475-489. https://doi.org/10.1080/19345747.2015.1017680

Archer, J., Kerr, K.A., & Pianta, R.C. (2014). Why measure effective teaching? In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 1-5). San Francisco, CA: John Wiley & Sons, Inc.

American Speech Language Hearing Association. (1993). *Definitions of communication disorders and variations*. Retrieved from http://www.asha.org/policy/RP1993-00208/

American Speech Language Hearing Association. (2004). *Evidence-based practice in communication disorders: Technical Report*. Retrieved from http://www.asha.org/policy/TR2004-00001/

American Speech Language Hearing Association. (2005). *Evidence-based practice in communication disorders*. Retrieved from http://www.asha.org/policy/PS2005-00221/

American Speech Language Hearing Association. (2010). *Roles and responsibilities of speech-language pathologists in schools*. Retrieved from http://www.asha.org/policy/PI2010-00317/

American Speech Language Hearing Association. (2014). *Performance assessment of contributions and effectiveness*. Retrieved from http://www.asha.org/Advocacy/state/PACE--Introduction/

American Speech Language Hearing Association. (2016). *Compendium of EBP guidelines and systematic reviews*. Retrieved from http://www.asha.org/Research/EBP/Compendium/

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R package version*, *1*(7), 1-23.

Berliner, D. C. (1987). Simple views of effective teaching and a simple theory of classroom instruction. In D. C. Berliner & B. Rosenshine (Eds.), *Talks to teachers* (pp. 93-110). New York: Random House.

Biancone, T. L., Farquharson, K., Justice, L. M., Schmitt, M. B., & Logan, J. A. (2014). Quality of language intervention provided to primary-grade students with language impairment. *Journal of Communication Disorders*, *49*, 13-24. https://doi.org/10.1016/j.jcomdis.2014.03.001

Black L.I, Vahratian A., Hoffman H.J. (2015). *Communication disorders and use of intervention services among children aged 3–17 years: United States, 2012*. Retrieved from the Centers for Disease Control and Prevention website: https://www.cdc.gov/nchs/data/databriefs/db205.htm

Blanton, L. P., Sindelar, P. T., & Correa, V. I. (2006). Models and measures of beginning teacher quality. *The Journal of Special Education*, *40*(2), 115-127. https://doi.org/10.1177/00224669060400020201

Blazar, D., Litke, E., & Barmore, J. (2016). What does it mean to be ranked a “high” or “low” value-added teacher? Observing differences in instructional quality across districts. *American Educational Research Journal*, *53*(2), 324-359. https://doi.org/10.3102/0002831216630407

Bui, X., Quirk, C., Almazan, S., & Valenti, M. (2010). *Inclusion education research and practice*. Retrieved from the Maryland Coalition on Inclusive Education website: http://www.mcie.org

Buckley, K., & Marion, S. (2011). *A survey of approaches used to evaluate educators in non-tested grades and subjects*. Retrieved from the National Center for the Improvement of Educational Assessment website: http://www.nciea.org/

Carnegie Classification of Institutions of Higher Education (2015). *Graduate Instructional Program Classification.* Retrieved from http://carnegieclassifications.iu.edu/

Center on Great Teachers and Leaders. (2014a). *Databases on state teacher and principal evaluation policies*. Retrieved from http://resource.tqsource.org/stateevaldb/

Center on Great Teachers and Leaders. (2014b). *Examples of state and district rubrics used to evaluate Specialized Instructional Support Personnel*. Retrieved from http://www.gtlcenter.org/

Corcoran, J. M. (2016). *Unannounced Classroom Observations: Impact on Teacher Practice and School Culture* (Doctoral dissertation). Retrieved from ProQuest Digital Dissertations.

Council for Academic Accreditation in Audiology and Speech-Language Pathology of the American Speech-Language-Hearing Association. (2017). 2017 Standards for Accreditation. Retrieved from https://caa.asha.org/reporting/standards/2017-standards/

Council for Clinical Certification in Audiology and Speech-Language Pathology of the American Speech-Language-Hearing Association. (2013). 2014 Standards for the Certificate of Clinical Competence in Speech-Language Pathology. Retrieved from http://www.asha.org/Certification/2014-Speech-Language-Pathology-Certification-Standards/

Council for Exceptional Children (2012). *CEC's position on special education teacher evaluation*. Retrieved from https://www.cec.sped.org/

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: ASCD.

Dollaghan, C. A. (2007). *The handbook for evidence-based practice in communication disorders*. Baltimore, MD: Paul H Brookes Publishing.

Dobbie, W., & Fryer Jr, R. G. (2013). Getting beneath the veil of effective schools: Evidence from New York City. *American Economic Journal: Applied Economics*, *5*(4), 28-60. https://doi.org/10.3386/w17632

Ebert, K. D. (2017). Measuring clinician–client relationships in speech-language treatment for school-age children. *American Journal of Speech-Language Pathology*, *26*(1), 146-152. https://doi.org/10.1044/2016_ajslp-16-0018

Ebert, K.D., & Kohnert, K. (2010). Common factors in speech-language treatment: Exploring qualities of effective clinicians. *Journal of Communication Disorders, 43,* 133-147. https://doi.org/10.1016/j.jcomdis.2009.12.002

Ebert, K. D., Kohnert, K., Pham, G., Disher, J. R., & Payesteh, B. (2014). Three treatments for bilingual children with primary language impairment: Examining cross-linguistic and cross-domain effects. *Journal of Speech, Language, and Hearing Research*, *57*(1), 172-186. https://doi.org/10.1044/1092-4388(2013/12-0388)

EduG (Version 6.1-e) [Computer software]. Swiss Society for Research in Education Working Group, Edumetrics Quality of Measurement in Education, Neuchatel, Switzerland.

Englert, C. S., Tarrant, K. L., & Mariage, T. V. (1992). Defining and redefining instructional practice in special education: Perspectives on good teaching. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, *15*(2), 62-86. https://doi.org/10.1177/088840649201500203

Evans, J. (1996). *Straightforward statistics for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole Publishing.

Every Student Succeeds Act. (2015). U.S. Department of Education. Retrieved from https://www.gpo.gov/

Family Guided Routines Intervention. (2008). Florida State University. Retrieved from http://fgrbi.fsu.edu/

Farquharson, K., Tambyraja, S. R., Logan, J., Justice, L. M., & Schmitt, M. B. (2015). Using hierarchical linear modeling to examine how individual SLPs differentially contribute to children's language and literacy gains in public schools. *American Journal of Speech-Language Pathology*, *24*(3), 504-516. https://doi.org/10.1044/2015_ajslp-14-0055

Fenstermacher, G. D. (1986). Philosophy of research on teaching: Three aspects. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 37-49). New York: Macmillan.

Fenstermacher, G. D., & Richardson, V. (2005). On making determinations of quality in teaching. *Teachers College Record*, *107*(1), 186-213. https://doi.org/10.1111/j.1467-9620.2005.00462.x

Ferguson, R. F., & Danielson, C. (2014). How framework for teaching and tripod 7Cs evidence distinguish key components of effective teaching. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 98-143). San Francisco, CA: John Wiley & Sons, Inc.

Ferguson, R. F., & Hirsch, E. (2014). How working conditions predict teaching quality and student outcomes. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 332-380). San Francisco, CA: John Wiley & Sons, Inc.

Gates Foundation. (2009). *Foundation commits $335 million to promote effective teaching and raise student achievement*. Retrieved from http://www.gatesfoundation.org/

Gates Foundation. (2013). *Measures of effective teaching project releases final research report.* Retrieved from http://www.gatesfoundation.org/

Gitomer, D., Bell, C., Qi, Y., McCaffrey, D., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record*, *116*(6), 1-32.

Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis*. Retrieved from The Center on Great Teachers & Leaders website: http://www.gtlcenter.org/

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Retrieved from The Center on Great Teachers & Leaders website: http://www.gtlcenter.org/

Goe, L., & Holdheide, L. (2011). *Measuring teachers' contributions to student learning growth for nontested grades and subjects*. Retrieved from The Center on Great Teachers & Leaders website: http://www.gtlcenter.org/

Gregory, A., Allen, J. P., Mikami, A. Y., Hafen, C. A., & Pianta, R. (2015). Eliminating the racial disparity in classroom exclusionary discipline. *Journal of Applied Research on Children: Informing Policy for Children at Risk*, *5*(2), 12. Retrieved from http://digitalcommons.library.tmc.edu/childrenatrisk/vol5/iss2/12/

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education*, *119*(3), 445-470. https://doi.org/10.1086/669901

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability*, *26*(1), 5-28. https://doi.org/10.1007/s11092-013-9179-5

Hancock, A. B., & Brundage, S. B. (2010). Formative feedback, rubrics, and assessment of professional competency through a speech-language pathology graduate program. *Journal of Allied Health*, *39*(2), 110-119. Retrieved from http://www.asahp.org/journal-of-allied-health/

Hancock, T. B., & Kaiser, A. P. (2006). Enhanced milieu teaching. In R. McCauley & M. Fey (Eds.), *Treatment of language disorders in children* (pp. 203–236). Baltimore, MD: Paul H. Brookes.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*(4), 430-511. https://doi.org/10.1080/07370000802177235

Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., Humez, A., & Lynch, K. (2012b). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment*, *17*(2-3), 88-106. https://doi.org/10.1080/10627197.2012.715019

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012a). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56-64. https://doi.org/10.3102/0013189x12437203

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Retrieved from the Bill & Melinda Gates Foundation website: http://k12education.gatesfoundation.org/

Holdheide, L. R., Goe, L.G., Croft, A., & Reschly, D. J. (2010). *Challenges in evaluating special education teachers and English Language Learner specialists*. Retrieved from The Center on Great Teachers & Leaders website: http://www.gtlcenter.org/

Holdheide, L. R., Hayes, L., & Goe, L.G. (2014). *Evaluating specialized instructional support personnel.* Retrieved from The Center on Great Teachers & Leaders website: http://www.gtlcenter.org/

Howell, D. C. (2007). *Statistical methods for psychology*. Belmont, CA: Thomson Higher Education.

IBM SPSS Statistics for Macintosh (Version 24.0) [Computer software]. Armonk, NY: IBM Corp.

Individuals with Disabilities Education Act. (2004). U.S. Department of Education. Retrieved from http://idea.ed.gov/

Joe, J. N., McClellan, C. A., & Holtzman, S. L. Scoring Design Decisions. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 415-443). San Francisco, CA: John Wiley & Sons, Inc.

Johnson, E. S. (2015). Increasing rural special education teacher candidates' ability to implement evidence-based practices: A program description of the Boise State University taters program. *Rural Special Education Quarterly*, *34*(1), 5. https://doi.org/10.1177/875687051503400103

Johnson, E. S., & Semmelroth, C. L. (2012). Examining interrater agreement analyses of a pilot special education observation tool. *Journal of Special Education Apprenticeship, 1*(2), 1-18. https://doi.org/10.1177/1534508413511489

Johnson, E., & Semmelroth, C. L. (2014). Special education teacher evaluation: Why it matters, what makes it challenging, and how to address these challenges. *Assessment for Effective Intervention, 39,* 71-82. https://doi.org/10.1177/1534508413513315

Johnson, S. M., Kraft, M. A., & Papay, J. P. (2012). How context matters in high-need schools: The effects of teachers' working conditions on their professional satisfaction and their students' achievement. *Teachers College Record*, *114*(10), 1-39. https://doi.org/10.1111/(issn)1467-9620

Jones, N. D., & Brownell, M. T. (2014). Examining the use of classroom observations in the evaluation of special education teachers. *Assessment for Effective Intervention*, *39*(2), 112-124. https://doi.org/10.1177/1534508413514103

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151. https://doi.org/10.1177/001316446002000116

Kamhi, A. G. (1995). Defining, developing, and maintaining clinical expertise: Research to practice. *Language, Speech, and Hearing Services in Schools*, *26*(4), 353-56. https://doi.org/10.1044/0161-1461.2604.353

Kane, T.J., Kerr, K., & Pianta, R. (Eds.). (2014). *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*. San Francisco, CA: John Wiley & Sons, Inc.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Retrieved from the Bill & Melinda Gates Foundation website: http://k12education.gatesfoundation.org/

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation*. Retrieved from the National Bureau of Economic Research website: http://www.nber.org/papers/w14607

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Retrieved

from the Bill & Melinda Gates Foundation website: http://k12education.gatesfoundation.org/

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of human Resources*, *46*(3), 587-613. https://doi.org/10.3386/w15803

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, *65*(23), 2276-2284. https://doi.org/10.2146/ajhp070364

Kohnert, K. (2013). *Language disorders in bilingual children and adults*. San Diego, CA: Plural Publishing.

Kraft, M. A., & Gilmour, A. F. (2016). Can principals promote teacher development as evaluators? A case study of principals' views and experiences. *Educational Administration Quarterly, 52*(5), 711-753. https://doi.org/10.1177/0013161x16653445

Kraft, M.A. & Gilmour, A.F. (2017). Revisiting the Widget Effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, *46*(5), 234-249.

Lash, A., Tran, L., and Huang, M. (2016). *Examining the validity of ratings from a classroom observation instrument for use in a district's teacher evaluation system (REL 2016–135)*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West. Retrieved from http://ies.ed.gov/ncee/edlabs.

Lawson, J. (2015). *Evaluating special education instructional practices using observation rubrics: Investigating the reliability of school administrator ratings* (Doctoral dissertation). Retrieved from http://escholarship.org/uc/item/37p867k8

Little, O., Goe, L., & Bell, C. (2009). *A practical guide to evaluating teacher effectiveness*. Retrieved from The Center on Great Teachers & Leaders website: http://www.gtlcenter.org/

Marzano, R. J., & Toth, M. D. (2013). *Teacher evaluation that makes a difference: A new model for teacher growth and student achievement*. Alexandria, VA: ASCD.

Maas, E., & Farinella, K. A. (2012). Random versus blocked practice in treatment for childhood apraxia of speech. *Journal of Speech, Language, and Hearing Research, 55*(2), 561-578. https://doi.org/10.1044/1092-4388(2011/11-0120)

Office of Special Education Programs. (2017). U.S. Department of Education. Retrieved from https://www2.ed.gov/about/offices/list/osers/osep/news-archive.html

Overholser, J. C. (2010). Clinical expertise: A preliminary attempt to clarify its core elements. *Journal of Contemporary Psychotherapy*, *40*(3), 131-139. https://doi.org/10.1007/s10879-009-9129-1

Papay, J. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*(1), 123-141. https://doi.org/10.17763/haer.82.1.v40p0833345w6384

Papay, J. P., & Johnson, S. M. (2012). Is PAR a good investment? Understanding the costs and benefits of teacher peer assistance and review programs. *Educational Policy*, *26*(5), 696-729. https://doi.org/10.1177/0895904811417584

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. (2016). *Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data*. Retrieved from the National Bureau of Economic Research website: http://www.nber.org/papers/w21986

Park, Y. S., Chen, J., & Holtzman, S. L. (2014). Evaluating efforts to minimize rater bias in scoring classroom observations. In T. J. Kane, K. A. Kerr, & R. C. Pianta

(Eds.), *Designing teacher evaluation systems* (pp. 381-414). San Francisco, CA: John Wiley & Sons, Inc.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109-119. https://doi.org/10.3102/0013189x09332374

Pianta, R. C., Hamre, B. K., Hayes, N., Mintz, S., & LaParo, K. M. (2008). *Classroom assessment scoring system–Secondary*. Charlottesville, VA: University of Virginia.

Pianta, R., & Kerr, K. A. (2014). Measuring Effective Teaching—The Future Starts Now. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 583-590). San Francisco, CA: John Wiley & Sons, Inc.

Polikoff, M. S. (2014). Does the Test Matter? Evaluating teachers when tests differ in their sensitivity to instruction. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 278-302). San Francisco, CA: John Wiley & Sons, Inc.

Praetorius, A. K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, *31*, 2-12. https://doi.org/10.1016/j.learninstruc.2013.12.002

Prince, C. D., Schuermann, P. J., Guthrie, J. W., Witham, P. J., Milanowski, A. T., & Thorn, C. A. (2006). *The other 69 percent: Fairly rewarding the performance of teachers of non-tested subjects and grades.* Retrieved from the Teachers & Leadership Programs website: https://www.tifcommunity.org

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rangamani, G., Coppens, P., Greenwald, M., & Keintz, C. (2016). Collaborative methods for training evidence-based practice: The triad model. *Contemporary Issues in Communication Science and Disorders*, *43*, 139-153. Retrieved from http://www.asha.org/Publications/cicsd/

Raudenbush, S. W., & Jean, M. (2014). To what extent do student perceptions of classroom quality predict teacher value added. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 170-202). San Francisco, CA: John Wiley & Sons, Inc.

Reinhorn, S. K., Johnson, S. M., & Simon, N. S. (2017). Investing in development: Six high-performing, high-poverty schools implement the Massachusetts teacher evaluation policy. *Educational Evaluation and Policy Analysis.* Advance online publication. https://doi.org/10.3102/0162373717690605

Rigby, J. G., Larbi-Cherif, A., Rosenquist, B. A., Sharpe, C. J., Cobb, P., & Smith, T. (2017). Administrator observation and feedback: Does it lead toward improvement in inquiry-oriented math instruction? *Educational Administration Quarterly, 53*(3), 475-516. https://doi.org/10.1177/0013161x16687006

Rivkin, S., Hanushek, E., & Kain, J. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458. Retrieved from https://www.econometricsociety.org/publications/econometrica/

Ruppar, A., Roberts, C., & Olson, A. J. (2015). Faculty perceptions of expertise among teachers of students with severe disabilities. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, *38*(3), 240-253. https://doi.org/10.1177/0888406414552331

Ruzek, E. A., Hafen, C. A., Hamre, B. K., & Pianta, R. C. (2014). Combining classroom observations and value added for the evaluation and professional development

of teachers. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 205-233). San Francisco, CA: John Wiley & Sons, Inc.

Rvachew, S., & Nowak, M. (2001). The effect of target-selection strategy on phonological learning. *Journal of Speech, Language, and Hearing Research*, *44*(3), 610-623. https://doi.org/10.1044/1092-4388(2001/050)

Sackett, D. L., Straus, S. E., Richardson, W. S., Rosenberg, W., & Haynes, R. B. (2000). *How to practice and teach EBM*. Edinburgh, Scotland: Churchill Livingstone.

Schmitt, M. B., Justice, L. M., & O'Connell, A. (2014). Vocabulary gain among children with language disorders: Contributions of children's behavior regulation and emotionally supportive environments. *American Journal of Speech-Language Pathology*, *23*(3), 373-384. https://doi.org/10.1044/2014_ajslp-12-0148

Schultz, S. E., & Pecheone, R. L. (2014). Assessing quality teaching in science. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 444-492). San Francisco, CA: John Wiley & Sons, Inc.

Semmelroth, C. L., & Johnson, E. (2014). Measuring rater reliability on a special education observation tool. *Assessment for Effective Intervention*, *39*(3), 131-145. https://doi.org/10.1177/1534508413511488

Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research*, *46*(4), 553-611. https://doi.org/10.2307/1169942

Sledge, A., & Pazey, B. L. (2013). Measuring teacher effectiveness through meaningful evaluation: Can reform models apply to general education and special education teachers? *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, *36*(3), 231-246. https://doi.org/10.1177/0888406413489839

Staiger, D. O. & Kane, T. J. (2014). Making decisions with imprecise performance measures: The relationship between annual student achievement gains and a

teacher's career value added. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 98-143). San Francisco, CA: John Wiley & Sons, Inc.

Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching project. *Education Finance and Policy, 10*(4), 535-572. https://doi.org/10.1162/edfp_a_00173

Stevens, J. P. (1992). *Applied multivariate statistics for the social sciences-2nd Edition*. Hillsdale, NJ: Erlbaum

Snyder, T.D., and Dillow, S.A. (2015). *Digest of education statistics* 2013 (NCES 2015-011). Retrieved from the National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education website: https://nces.ed.gov/programs/digest/

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *The American Economic Review*, *102*(7), 3628-3651. https://doi.org/10.1257/aer.102.7.3628

Teachstone (2013). *The CLASS tool*. Charlottesville, VA: Teachstone.

Thomas, S. (2001). Dimensions of secondary school effectiveness: Comparative analyses across regions. *School Effectiveness and School Improvement*, *12*(3), 285-322. https://doi.org/10.1076/sesi.12.3.285.3448

Ukrainetz, T. A. (2014). *School-age language intervention: Evidence-based practices*. Austin, TX: Pro-Ed, Inc.

van de Grift, W. J. C. M. (2014). Measuring teaching quality in several European countries. *School effectiveness and school improvement*, *25*(3), 295–311. https://doi.org/10.1080/09243453.2013.794845

Van de Grift, W., & Van der Wal, M. (2010). Measuring the development of professional competence among teachers. Retrieved from http://www.icsei.net/icsei2011/Full%20Papers/0127_A.pdf.

Vickers, S. T. (2015). *The extent to which annual professional performance reviews change classroom instructional practice: A sequential mixed-methods study of teacher evaluations in central New York* (Doctoral dissertation). Retrieved from http://fisherpub.sjfc.edu/education_etd/242/

Walkington, C., & Marder, M. (2013). Classroom observation and value-added models give complementary information about quality of mathematics teaching. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 234-277). San Francisco, CA: John Wiley & Sons, Inc.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Retrieved from The New Teacher Project website: https://tntp.org

Williams, A. L., McLeod, S., & McCauley, R. J. (2010). *Interventions for speech sound disorders in children*. Baltimore, MD: Brookes Publishing Company.

Winter, B. (2013). *A very basic tutorial for performing linear mixed effects analyses*. Retrieved from: http://www.bodowinter.com/tutorial/bw_LME_tutorial2.pdf

Appendix A

2014-2015 Performance Evaluation Instrument

DOMAIN 1: Planning of speech/language service				
	Requires Attention	**Developing**	**Proficient**	**Exemplary**
Element 1: Uses knowledge of Evidence-Based Practice (EBP) to select learning targets and plan service	*SLP does NOT demonstrate familiarity with EBP when selecting learning targets or planning implementation of service.*	*SLP demonstrates some familiarity with EBP when selecting learning targets and planning implementation of service.*	*SLP applies EBP when selecting learning targets and planning implementation of service.*	*SLP demonstrates extensive knowledge of EBP when selecting learning targets and planning implementation of service.*
LOOK FORs (not an exhaustive list):	SLP cannot explain the EBP rationale for the selection of learning targets and planning of service implementation.	SLP struggles to explain the EBP rationale for the selection of learning targets and planning of service implementation, but is aware of EBP resources within or outside of the district.	SLP demonstrates the EBP process in the planning of service.	SLP models the EBP process in the planning of service (e.g., professional development for SLPs or staff; explicit sharing of EBP references with others).
Evidence of Performance:				
Element 2: Designs coherent, sequential speech/language sessions that are aligned with student needs (e.g., IEP goals/objectives)	*Planned learning activities and instructional strategies are NOT aligned with student needs.*	*Planned learning activities and instructional strategies are aligned with student needs.*	*Planned learning activities and instructional strategies are aligned with student needs, sequential (build on prior student knowledge or performance), and facilitate further student construction of knowledge.*	*Planned learning activities and instructional strategies are aligned with student needs, sequential (build on prior student knowledge or performance), facilitate further construction of student knowledge, and are linked to learning activities in other instructional areas.*
LOOK FORs (not an exhaustive list):	1) SLP does not have a plan; **OR** 2) SLP has a plan that is not aligned with IEP goals/objectives.	Activities are planned and materials are prepped based on IEP goals/objectives, but the plan is not sequential.	Activities are planned and materials are prepped based on IEP goals/objectives; and the plan is sequential.	1) Activities are planned and materials are prepped based on IEP goals/objectives; and the plan is sequential. **AND** 2) Planning explicitly addresses transfer of communication skills to other instructional areas (e.g., collaboration meetings to discuss transfer).
Evidence of Performance:				
Element 3: Designs speech/language sessions that place learning targets in functional contexts ("real-life", applicable to everyday experiences)	*Planning indicates the SLP has NOT collaborated with the student's communication partners to determine functional contexts.*	*Planning indicates the SLP has collaborated with the student's communication partners to determine functional contexts.*	*Planning indicates the SLP has collaborated with the student's communication partners to determine functional contexts and has extensive knowledge of the student's communication settings and associated needs.* *Learning targets are linked to a generalization plan across two functional settings/routines (e.g., classrooms, specialists, lunchroom, playground, home, community settings, etc.).*	*Planning indicates the SLP has collaborated with the student's communication partners to determine functional contexts and has extensive knowledge of the student's communication settings and associated needs.* *Learning targets are linked to a generalization plan across at least three functional settings/routines (e.g., classrooms, specialists, lunchroom, playground, home, community).*

LOOK FORs (not an exhaustive list):	Planning shows: 1) No evidence of push-in service and/or collaboration with student's team. 2) No real-world communication situations or opportunities for independent communication.	Planning shows: 1) Some evidence of push-in service and/or collaboration with student's team, but learning activities and instructional strategies are primarily generated from materials that are not authentic forms of communication; **AND** 2) Some real-world communication situations and opportunities for independent communication.	Planning shows: 1) Clear evidence of push-in service and/or collaboration with student's team; **AND** 2) Learning activities and instructional strategies represent authentic forms of communication; **AND** 3) Many real-world communication situations and opportunities for independent communication.	Planning shows: 1) Clear evidence of push-in service and/or collaboration with student's team; **AND** 2) Learning activities and instructional strategies represent authentic forms of communication; **AND** 3) Many real-world communication opportunities for independent communication; **AND** 4) Explicit ideas for generalization of skills.
Evidence of Performance:				
Element 4: Designs speech/language sessions that are culturally relevant and personally meaningful	*Planned learning activities and instructional strategies are NOT informed by knowledge of student's varied skills, interests, and/or cultural/linguistic backgrounds.*	*Planned learning activities and instructional strategies address one or more aspects of student's varied skills, interests, and/or cultural/linguistic backgrounds, but some assumptions (i.e., stereotypes, over-generalizations) have been made about students.*	*Planned learning activities and instructional strategies are informed by knowledge of student's varied skills, interests, and/or cultural/linguistic backgrounds and avoid assumptions (i.e., stereotypes, over-generalizations) about students.*	*SLP models the design of learning activities/instructional strategies that are culturally relevant and personally meaningful through mentorship, coaching, and/or assistance to others.*
LOOK FORs (not an exhaustive list):	Planning shows: 1) Materials do not reflect the student's varied skills, interests, or cultural/linguistic backgrounds; **AND/OR** 2) Assumptions (i.e., stereotypes, overgeneralizations) are present.	Planning shows: Materials reflect student's varied skills, interests, and/or cultural/linguistic backgrounds, but assumptions (i.e., stereotypes, overgeneralizations) are present.	Planning shows: 1) Materials reflect student's varied skills, interests, and varied cultural/linguistic backgrounds; **AND** 2) Assumptions (i.e., stereotypes, overgeneralizations) are not present; **AND** 3) SLP has connected with family members and/or cultural liaisons to better understand the student's individual cultural/linguistic background.	1) SLP develops cultural and/or personally meaningful resources to aid others in the planning of service; **AND** 2) SLP has a system for updating his/her knowledge about service delivery to students who are culturally/linguistically diverse (e.g., professional development, regularly connecting with family/cultural liaisons).
Evidence of Performance:				
Element 5: Plans for relevant resources	*Plans do NOT include the use of relevant resources to present information.*	*Plans include the use of relevant resources to present information.*	*Plans include the use of relevant resources to promote increased cognitive engagement and deeper conceptual understanding.*	*Plans include the use of multiple resources to promote increased cognitive engagement and deeper conceptual understanding.* *Students have opportunities to use relevant resources to enrich their understanding of the content.*
LOOK FORs (not an exhaustive list):	Resources may include any or all of the following: ● Therapy materials (games, flashcards, toys, mirrors, etc.). ● Books ● Visual aids ● Routine-based materials in students' homes ● iPads and/or tech ● AAC	Resources may include any or all of the following: ● Therapy materials (games, flashcards, toys, mirrors, etc.). ● Books ● Visual aids ● Routine-based materials in students' homes ● iPads and/or tech ● AAC	Resources may include any or all of the following: ● Therapy materials (games, flashcards, toys, mirrors, etc.). ● Books ● Visual aids ● Routine-based materials in students' homes ● iPads and/or tech ● AAC	Resources may include any or all of the following: ● Therapy materials (games, flashcards, toys, mirrors, etc.). ● Books ● Visual aids ● Routine-based materials in students' homes ● iPads and/or tech ● AAC
Evidence of Performance:				

Element 6: Plans for assessment strategies to monitor student progress	*Plans do NOT include the use of assessment strategies to monitor student progress and adapt instruction to student needs.*	*Plans include limited use of assessment strategies to monitor student progress and adapt instruction to student needs.*	*Plans include frequent use of assessment strategies to monitor student progress and adapt instruction to student needs.*	*Plans include frequent use of assessment strategies to monitor student progress and adapt instruction to student needs.* *Planned assessments are designed within an EBP framework.*
LOOK FORs (not an exhaustive list):	SLP's response in pre-conference indicates a lack of understanding of the role of assessment.	Session plan includes limited opportunities to assess student's learning.	Session plan includes sufficient opportunities to assess student's learning.	1) Session plan includes sufficient opportunities to assess student's learning; **AND** 2) Plans include references to assessments in EBP literature; **AND/OR** 3) SLP is planning an action research project in order to compare collected data with data in external literature.
Evidence of Performance:				
DOMAIN 2: Climate of speech/language service				
	Requires Attention	**Developing**	**Proficient**	**Exemplary**
Element 1: Establishes high expectations for student participation and explicitly sets up the environment so students understand the schedule and purpose of the service session	*Task instructions and learning interactions do NOT include expectations for student participation.*	*Most learning activities are characterized by passive student participation; **OR*** *The schedule and purpose of service events is not explicit.*	*Learning activities are characterized by active student participation; **AND*** *The schedule and purpose of service events is explicit (e.g., "Today, we are going to…because…").*	*Learning activities are characterized by active student participation; **AND*** *The schedule and purpose of service events is explicit (e.g., "Today, we are going to…because..."); **AND*** *SLP encourages students to take on new challenges (e.g., "Let's try this") while still ensuring student success.*
LOOK FORs (not an exhaustive list):	Student is not made aware of participation expectations.	1) SLP does not encourage active participation; **AND/OR** 2) SLP does not explicitly state plan and purpose of the lesson.	1) SLP makes clear student is expected to participate; **AND** 2) SLP explicitly states plan and purpose of the lesson.	1) SLP makes clear student is expected to participate; **AND** 2) SLP explicitly states plan and purpose of the lesson; **AND** 3) Challenges are explicit.
Evidence of Performance:				
Element 2: Establishes, maintains, and explicitly communicates room/space routines and procedures (with respect to your service delivery model/setting)	*Room/space routines and procedures are NOT explicitly communicated; excessive instructional time is lost due to lack of room/space routines (with respect to your service delivery model/setting).*	*Room/space routines and procedures are generally communicated; some instructional time is lost due to lack of room/space routines (with respect to your service delivery model/setting).*	*Room/space routines and procedures are explicitly communicated; minimal instructional time is lost due to lack of room/space routines (with respect to your service delivery model/setting).*	*Room/space routines and procedures are explicitly communicated; minimal instructional time is lost due to lack of room/space routines (with respect to your service delivery model/setting). Transitions are organized and orderly.*
LOOK FORs (not an exhaustive list):	Student is not made aware of the room routines and procedures for the speech/language service.	1) SLP gives warnings for transitions when appropriate; **AND/OR** 2) SLP maintains a room/space routine as a framework for learning; **AND/OR** 3) Some learning time is lost due to lack of room/space routines.	1) SLP gives warnings for transitions when appropriate; **AND/OR** 2) SLP maintains a room/space routine as a framework for learning; **AND/OR** 3) Minimal learning time is lost due to lack of room/space routines.	1) SLP gives warnings for transitions when appropriate; **AND** 2) SLP maintains a room/space routine as a framework for learning; **AND** 3) Minimal learning time is lost due to lack of room/space routines.
Evidence of Performance:				

Element 3: Uses effective and constructive behavior management (refers to social/emotional behavior; "comportment")	*SLP does NOT provide constructive or timely feedback about student behavior/comportment.*	*SLP occasionally provides behavior/comportment feedback that is reactive, but not preventative.*	*SLP provides behavior/comportment feedback that is constructive, timely, and preventative.*	*SLP provides behavior/comportment feedback that is constructive, timely, and preventative.* *Students receive acknowledgment of expected behavior/comportment, as appropriate.*
LOOK FORs (not an exhaustive list):	1) SLP ignores both positive and negative student behaviors; **AND/OR** 2) SLP does not provide behavior/comportment feedback that is constructive.	1) Feedback immediately follows behavior, but is not preventative. 2) SLP redirects student when appropriate.	1) Feedback is constructive, immediately follows behavior, and prevents negative behavior from occurring. 2) SLP redirects student when appropriate.	1) Feedback is constructive, immediately follows behavior, and prevents negative behavior from occurring; **AND** 2) SLP redirects student when appropriate; **AND** 3) SLP's acknowledgement of expected student behavior is appropriate.
Evidence of Performance:				
Element 4: SLP builds positive relationships (i.e., rapport) with students	*SLP does NOT demonstrate a positive regard towards students or talks negatively about other students.*	*SLP demonstrates a positive regard towards students, allows students to make mistakes, pays students compliments, and intervenes when students need support.*	*SLP demonstrates a positive regard towards students, allows students to make mistakes, pays students compliments, intervenes when students need support, ensures relaxed/comfortable interactions, and offers students choices when appropriate.*	*SLP demonstrates a positive regard towards students, allows students to make mistakes, pays students compliments, intervenes when students need support, ensures relaxed/comfortable interactions, offers students choices when appropriate, and makes no power-affirming remarks (e.g., "because I said so") or actions.*
LOOK FORs:	N/A	N/A	N/A	N/A
Evidence of Performance:				
DOMAIN 3: Implementation of Speech/Language Service				
	Requires Attention	**Developing**	**Proficient**	**Exemplary**
Element 1: Explicitly communicates speech/language learning objectives	*Learning objectives are NOT communicated.*	*For each activity, learning objectives are generally communicated and clarified for the student.*	*For each activity, learning objectives are explicitly communicated and clarified for the student.*	*For each activity, learning objectives are explicitly communicated and clarified for the student.* *SLP has established systems for explicitly communicating learning objectives and makes individual adjustments as needed.*
LOOK FORs (not an exhaustive list):	N/A	N/A	N/A	1) Systems may include table tent cards, posters, chart, notebook, or other visual/tactile systems 2) Adjustments and/or adaptations in communication occur (e.g., not talking at too high/low level for the student's needs).
Evidence of Performance:				

Element 2: Provides learning activities and uses instructional strategies that are engaging and motivating	*SLP does NOT implement learning activities and instructional strategies that engage or motivate students.*	*Few of the selected learning activities and instructional strategies engage and motivate students.*	*Many of the selected learning activities and instructional strategies engage and motivate students.*	*SLP implements many learning activities and instructional strategies that engage and motivate students in ways that are culturally relevant and/or personally meaningful.*
LOOK FORs (not an exhaustive list):	Student has a negative response to learning activities.	1) Student does not react negatively to learning activities; **AND/OR** 2) Student occasionally disengages with learning activities.	1) Student rarely disengages with learning activities; **AND/OR** 2) Student may request to continue learning activities or use of learning materials (books, games, technology).	1) Statements and questions connect with student experience; **AND** 2) Materials are culturally relevant and/or personally meaningful to student; **AND** 3) SLP acknowledges student is proud of his/her learning or effort.
Evidence of Performance:				
Element3: Provides opportunities to practice skills and demonstrate learning at a rigorous pace while ensuring student success	*Students do NOT have opportunities to practice or demonstrate what they are learning.*	*Students have some opportunities to practice or demonstrate what they are learning.* *Pacing is either too fast or slow for the student.*	*Students have many opportunities to practice or demonstrate what they are learning.* *Pacing is rigorous and ensures student success.*	*Students have many opportunities to practice or demonstrate what they are learning.* *Pacing is rigorous and ensures student success.* *SLP connects practice opportunities to what students are learning in other classroom and/or home environments.*
LOOK FORs (not an exhaustive list):	SLP presents information, only, with no student opportunities to practice a skill before moving to the next skill.	SLP uses less than 50% of the available opportunities for student practice.	SLP uses at least 50% of the available opportunities for student practice.	1) SLP uses at least 50% of the available opportunities for student practice; **AND** 2) Practice opportunities clearly connect to classroom/home environments (e.g., using words from classes, textbooks, or home).
Evidence of Performance:				
Element 4: Uses instructional techniques to promote cognitive engagement with learning targets	*SLP does NOT use instructional techniques to cognitively engage students; most learning activities are directed below the student's developmental level.*	*SLP uses instructional techniques to cognitively engage students but most learning activities are directed at the student's developmental level.SLP occasionally provides effective wait time.*	*SLP uses instructional techniques to cognitively engage students in learning activities directed at and above the student's developmental level.SLP frequently provides effective wait time.*	*SLP uses instructional techniques to cognitively engage students in learning activities directed at and above the student's developmental level.SLP frequently provides effective wait time.* *SLP acknowledges student effort for engaging with learning activities above the student's developmental level.*
LOOK FORs (not an exhaustive list):	SLP does not challenge students.	1) Questions and discussions provide some challenge to students; **AND** 2) There are some opportunities for independent problem-solving.	1) Questions and discussions provide challenge to students; **AND** 2) There are frequent opportunities for independent problem-solving.	1) SLP consistently develops learning experiences where independent problem-solving is valued; **AND** 2) SLP acknowledges effort towards and provides opportunities for self-monitoring of independent problem-solving.
Evidence of Performance:				

Element 5: Uses assessment strategies to monitor student progress	*SLP does NOT monitor student progress through formal data collection and/or informal assessment.*	*SLP monitors student progress through formal data collection and/or informal assessment, but infrequently adjusts instruction based on student responses (e.g., increasing or decreasing task difficulty).*	*SLP monitors student progress through formal data collection and/or informal assessment and frequently adjusts instruction based on student responses (e.g., increasing or decreasing task difficulty).*	*SLP monitors student progress through formal data collection and/or informal assessment, frequently adjusts instruction based on student responses (e.g., increasing or decreasing task difficulty), and interprets assessment data and/or adjusted instruction within an EBP framework.*
LOOK FORs (not an exhaustive list):	SLP does not use assessment strategies.	SLP monitors performance and adjustments are made in less than 50% of opportunities.	SLP monitors performance and adjustments are made in at least 50% of opportunities.	SLP uses data to make adjustments based on external EBP literature (or, other levels of EBP evidence).
Evidence of Performance:				
Element 6: Gives explicit and timely feedback to improve student learning	*Students do NOT receive feedback to promote learning.*	*The majority of feedback students receive is positive, but not explicit to learning targets.*	*Feedback to students is explicit, timely, and based on individual student characteristics.*	*Feedback to students is explicit, timely, based on individual student characteristics, and part of a scaffolding plan to reduce feedback dependence.*
LOOK FORs (not an exhaustive list):	Feedback is not provided.	Feedback is not specific to learning.	Feedback is specific and guides learning.	Scaffolding plan is evident.
Evidence of Performance:				
DOMAIN 4: Professional Responsibilities, Due Process Documentation, and Case Management				
	Requires Attention	**Developing**	**Proficient**	**Exemplary**
Element 1: Uses self-reflection and self-assessment to improve service to students	*SLP does NOT show evidence of self-reflection and self-assessment.*	*SLP reflects on student data/work to identify effectiveness of service to students.*	*SLP reflects on his or her own effectiveness using artifacts and student data/work to identify areas of strength and areas for growth.*	*SLP reflects on his or her own effectiveness using artifacts and student work to identify areas of strength and areas for growth.* *SLP models reflective practices for other staff through peer coaching, sharing, facilitating, and/or modeling in professional learning communities.*
LOOK FORs (not an exhaustive list):	No evidence of self-reflection.	N/A	SLP reflects on any of the following: 1) Parent or student surveys 2) Parent or staff contact 3) Student data	SLP models reflective practices for colleagues by developing tools and resources to enhance peer coaching or self-assessment.
Evidence of Performance:				
Element 2: Uses feedback to improve service to students	*SLP does NOT seek or use feedback from colleagues, administrators, families, and students to improve service to students.*	*SLP seeks feedback from colleagues, administrators, families, and students to enhance professional practice.*	*SLP seeks feedback from colleagues, administrators, families, and students to enhance professional practice.* *SLP uses feedback to implement instructional approaches that result in increased student learning.* *SLP has systems in place for obtaining feedback from colleagues/ administrators/ families/ students.*	*SLP seeks feedback from colleagues, administrators, families, and students to enhance professional practice.* *SLP uses feedback to implement instructional approaches that result in increased student learning.* *SLP has systems in place for obtaining feedback from colleagues/ administrators/families/ students.* *SLP develops and shares feedback resources with colleagues.*

LOOK FORs (not an exhaustive list):	No evidence SLP seeks feedback, or, SLP rejects feedback from others.	Seeks feedback based on any of the following: 1) PLC logs (e.g., difficult cases) 2) Parent and/or student surveys. 3) Parent and/or staff contact regarding student needs or performance. 4) Student data.	Seeks and uses feedback based on any of the following: 1) PLC logs (e.g., difficult cases) 2) Parent and/or student surveys. 3) Parent and/or staff contact regarding student needs or performance. 4) Student data.	N/A
Evidence of Performance:				
Element 3: Participates in relevant professional development (PD)	*SLP does NOT participate in professional learning/PD activities designed to improve speech/language service.*	*SLP participates in professional learning/PD activities designed to improve speech/language service, but there is limited implementation of professional learning/PD activities.*	*SLP participates in and implements professional learning/PD activities designed to improve speech/language service.*	*SLP provides professional learning/PD activities designed to improve speech/language service.*
LOOK FORs (not an exhaustive list):	No evidence of professional development (e.g., no PLC attendance; no CEUs).	Evidence of professional learning: 1. CEUs 2. PLC activities	Evidence of professional learning: 1. CEUs 2. PLC activities CEUs Implementation: Student work samples.	Provision of professional learning for others. "Others" may include SLPs, teachers, SEAs, administrators, etc.
Evidence of Performance:				
Element 4: Collaborates with educational team	*Staff Survey*	*Staff Survey*	*Staff Survey*	*Staff Survey*
LOOK FORs:	N/A	N/A	N/A	N/A
Evidence of Performance:				
Element 5: Collaborates with families	*SLP does NOT communicate and/or build relationships with families.*	*SLP responds to family requests for communication regarding student progress.*	*SLP responds to and initiates family requests for communication, conducts interactions with families that result in positive relationships, displays sensitivity for families and involves them in problem solving.*	*SLP responds to and initiates family requests for communication, conducts interactions with families that result in positive relationships, displays sensitivity for families and involves them in problem solving.* *SLP models family communication by setting up ongoing communication systems.*
LOOK FORs (not an exhaustive list):	SLP does not respond to parent requests for contact.	When contacted by families, SLP responds as evidenced by: 1) Due process contact logs; 2) Progress notes; or 3) Phone calls, conversations, etc.	SLP initiates contact as evidenced by: 1) Due process contact; 2) Progress notes; or 3) Calls, conversations. SLP seeks out information and uses family liaisons to increase collaboration with families. SLP avoids professional jargon and makes communication adjustments to meet the needs of families.	SLP systems might include: 1) Communication notebooks 2) A dedicated time for contacting families/caregivers 3) Coordination of care with outside providers 4) Home interventions/ practice
Evidence of Performance:				

Element 6: Applies knowledge of the pre-referral (RTI) process in the prevention and identification of disabilities **Pre-referral in Birth-5 may include parent strategies*	*SLP does NOT demonstrate knowledge of pre-referral procedures.*	*SLP demonstrates knowledge of pre-referral procedures.*	*SLP demonstrates knowledge of pre-referral procedures, problem-solves pre-referral implementation, and collaborates with educational staff to interpret pre-referral data.*	*SLP models pre-referral processes by organizing an ongoing pre-referral system and/or through mentorship, coaching, and/or offering assistance to others.*
LOOK FORs (not an exhaustive list):	SLP is unable to explain the essential steps of pre-referral intervention: teacher concern, team meeting to design intervention, intervention implementation, and data analysis.	SLP is able to explain the essential steps of pre-referral intervention: teacher concern, team meeting to design intervention, intervention implementation, and data analysis.	1) SLP is able to explain the essential steps of pre-referral intervention and discuss associated problem-solving; **AND** 2) SLP shows evidence of pre-referral collaboration.	1) SLP develops and models the use of pre-referral resources; **AND** 2) SLP has a system for updating pre-referral knowledge; **AND** 3) SLP has set-up a system to address pre-referral at his/her site.
Evidence of Performance:				
Element 7: Provides evaluations that are appropriate, accurate, and educationally-focused	*SLP does NOT provide evaluations that are appropriate, accurate, or educationally-focused.*	*SLP selects appropriate standardized and non-standardized assessment measures for the area(s) of presenting concern and individual student needs and accurately administers and scores the assessment measures; addresses state Eligibility Criteria accurately; SLP does not, however, consistently interpret all assessment data with a focus on educational relevance.* *For children who are culturally/linguistically diverse, consideration of their cultural/linguistic background is explicit in the evaluation report.*	*SLP selects appropriate standardized and non-standardized assessment measures for the area(s) of presenting concern and individual student needs; accurately administers and scores the assessment measures; addresses state Eligibility Criteria accurately; interprets all assessment data with a focus on educational relevance; and writes the evaluation report in family-accessible language.* *For children who are culturally/linguistically diverse, consideration of their cultural/linguistic background is explicit in the evaluation report.*	*SLP selects appropriate standardized and non-standardized assessment measures for the area(s) of presenting concern and individual student needs; accurately administers and scores the assessment measures; addresses state Eligibility Criteria accurately; interprets all assessment data with a focus on educational relevance; writes the evaluation report in family-accessible language; and the Evaluation Educational Needs Statement is consistent with the analysis of the evaluation measures.* *For children who are culturally/linguistically diverse, consideration of their cultural/linguistic background is explicit in the evaluation report.*
LOOK FORs:	N/A	N/A	N/A	N/A
Evidence of Performance:				
Element 8: Proposes educational plans that are complete, educationally relevant, and measurable ** Educational plans may include an IFSP, ISP, or IEP.*	*SLP does NOT provide a service plan that is complete, educationally relevant, or measurable*	*SLP provides a service plan that is complete but not necessarily educationally relevant nor consistently measurable.* *For children who are culturally/linguistically diverse, consideration of their cultural/linguistic background is explicit in the service plan.*	*SLP provides a service plan that is complete, educationally relevant, measurable, and goal(s) and objectives are supported by data explicitly stated in the Present Levels of Performance.* *For children who are culturally/linguistically diverse, consideration of their cultural/linguistic background is explicit in the service plan.*	*SLP provides a service plan that is complete, educationally relevant, measurable, and goal(s) and objectives are supported by data explicitly stated in the Present Levels of Performance.* *Service plan is written in family-accessible language.* *For children who are culturally/linguistically diverse, consideration of their cultural/linguistic background is explicit in the service plan.*
LOOK FORs:	N/A	N/A	N/A	N/A
Evidence of Performance:				

Element 9: Provides a continuum of service in the least restrictive environment	*SLP does NOT design and implement a continuum of service in the LRE.*	*SLP occasionally designs and implements a continuum of service in the LRE.*	*SLP frequently designs and implements a continuum of service in the LRE.*	*SLP models the design and implementation of a continuum of service through mentorship, coaching, and/or offering assistance to others.*
LOOK FORs (not an exhaustive list):	1) LRE statement of rationale is not present in service plan (e.g., IEPs); **AND** 2) SLP's schedule shows NO variation in services; **OR** 3) SLP does NOT adjust service model or consider dismissal, even when data suggest otherwise; **OR** 4) SLP is NOT familiar with the continuum of service options.	1) LRE statement of rationale is present in service plan (e.g., IEPs); **AND** 2) SLP's schedule shows some variation in service delivery, frequency and location; **AND/OR** 3) Minimal adjustment of service model (including dismissal), even when data suggest otherwise.	1) LRE statement of rationale is present in service plan (e.g., IEPs); **AND** 2) SLP's schedule shows variation in service delivery, frequency and location; **AND/OR** 3) Adjustment of service model (including dismissal), when appropriate.	1) LRE statement of rationale is present in service plan (e.g., IEPs); **AND** 2) SLP extensively collaborates with educational staff to determine service continua; **AND** 3) SLP has extensive knowledge of the benefits of LRE; **AND** 4) SLP models a continuum of service for colleagues, staff, SEAs, administrators, caregivers, practicum students, etc.
Evidence of Performance:				
Element 10: Establishes data collection systems for all provided services	*Across the provided services, SLP does NOT have established systems for gathering, recording, and summarizing data.*	*SLP has an established system for gathering, recording, and summarizing data for most of the services provided.*	*SLP has an established system for gathering, recording, and summarizing data for each service provided.*	*SLP has an established system for gathering, recording, and summarizing data for each service based on EBP.*
LOOK FORs	N/A	N/A	N/A	N/A
Evidence of Performance:				
Element 11: Makes data-based decisions (DBD) for all provided services	*SLP does NOT make DBD for instructional purposes.*	*SLP makes DBD for some of the services provided including but not limited to:* *- Assessment and diagnosis (e.g., interpretation of data measures)* *- Intervention (e.g., goals/objectives, scaffolding support, selection of materials/activities.)* *- LRE (e.g., dismissal)*	*SLP makes DBD for each of the services provided including but not limited to:* *- Assessment and diagnosis (e.g., interpretation of data measures)* *- Intervention (e.g., goals/objectives, scaffolding support, selection of materials/activities.)* *- LRE (e.g., dismissal)*	*SLP makes DBD for each of the services provided including but not limited to:* *- Assessment and diagnosis (e.g., interpretation of data measures)* *- Intervention (e.g., goals/objectives, scaffolding support, selection of materials/activities.)* *- LRE (e.g., dismissal)* *SLP interprets data within an EBP framework.*
LOOK FORs	N/A	N/A	N/A	SLP compares his/her collected data with data found in external literature (or, other EBP evidence) to make DBD for provided services.
Evidence of Performance:				

Appendix B

Collaborating Staff Survey in the 2014-2015 School Year

Dear Colleagues,
Part of the observation (evaluation) process for speech-language pathologists (SLPs) includes gathering feedback on their ability to collaborate with other staff members at their buildings. As a colleague, you have been invited to complete this survey, along with several other colleagues at your building.

On the following pages, you will see a short survey of questions. This survey is ANONYMOUS and voluntary. The estimated time to complete this survey is 4-6 minutes.

Thank you for your participation. We deeply appreciate your time!

Name of collaborating SLP (for whom you were invited to provide feedback for on this survey):

---

1. The SLP responds to requests for communication in a reasonable amount of time:
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree
2. The SLP limits the amount of technical terms and when used, explains them effectively to team members:
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree
3. The SLP is open to feedback from team members:
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree
4. The SLP demonstrates effective problem-solving skills:
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree
5. The SLP collaborates with others to plan services:
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree
6. The SLP is a knowledgeable resource about speech-language disabilities:
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree
7. The SLP provides specific insights about students with speech-language disabilities:
   - ❑ Strongly Agree
   - ❑ Agree

- ❑ Disagree
- ❑ Strongly Disagree

8. The SLP demonstrates leadership skills at my school/site:
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree

Additional items for SLPs only:

9. The SLP provides accurate information about his/her caseload to ensure caseload equity within the school/site:
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree
10. The SLP attends agreed-upon meetings with other SLPs at the school/site:
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree
11. The SLP provides smooth transitions for students/families within the school/site by accurately completing all due process documentation:
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree
12. The SLP is flexible in the division of caseload and willing to work across grades/disabilities:
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree
13. The SLP is flexible in the use of space/materials:
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree
14. The SLP is flexible in the division of school-wide responsibilities (e.g., pre-referral intervention meetings; special education team meetings; grade-level team meetings):
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree
15. The SLP provides support to other SLPs at the school/site:
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree

Appendix C

Example of Completed Pre-Observation Questions by an SLP in 2014-2015 Evaluations

Domain 1 Questions (Planning of Service)

Dear SLP,

Please complete the form below with brief notes (feel free to use bullet points, short notes, etc.). In general, we'll spend approximately 2-3 minutes discussing each question during our pre-observation conference.

Thank you!

*The term "student" in the questions below may include an individual student, a group of students, or family/child.

1. Describe if/how Evidence-Based Practice (EBP) influenced your decision of learning targets and/or the structure of today's lesson plan:

Evidence-Based Practice has influenced my decision making of today's learning targets. When I initially started providing services for this student, I used the Minimal Pair Approach (Baker, 2010). After experimenting with the Minimal Pair Approach, I noticed that he was having difficulty discriminating between the minimally different sounds. After attending the "EBP for Speech Sound Disorders" training earlier this year, I decided to explore the Maximal Oppositions Approach (Gierut 1989; Gierut 1990; Gierut, 1992) with him to see if this intervention would be more successful. The EBP for Maximal Oppositions is as follows:

-Scientific Evidence: Gierut 1989; Gierut 1990; Gierut,1992. Because this student has a moderate to severe phonological disorder, research evidence supports that this intervention may have a better outcome than Minimal Pairs.

-Clinical Expertise: I have been using this approach with this student for a couple of months and have found it be successful in improving his production of the /g/ sound. I am able to plan the intervention targets well.

-Client Perspectives: The student seemed to enjoy the activities/games that I have used with this approach. Also, he has displayed positive attitudes towards the success he has made with this approach. It is much easier for him to discriminate and produce maximally different sounds.

Additionally, describe any modeling of the EBP above. For example, have you modeled the EBP process for above for your peers and/or lead any EBP trainings for above? Feel free to write "N/A" here:

N/A

2. Describe the general structure/schedule of today's lesson plan and when applicable, how today's lesson plan builds on previous and/or future lesson plans.

We have been using the Maximal Pair Approach to target the production of the "ch" sound in the initial position of words in previous sessions. This is a targeted sound from his IEP. In future lesson, I hope to target "ch" in the final position of words. The general structure for today's lesson will be as follows:

-Review visual schedule of events

-Review targeted sounds ("m" and "ch")

-Alphabeats reading and writing of targeted sounds ("m" and "ch")

-Demonstration of making the "ch" and "m" sounds (use iPad and/or mirror)

-Discriminate between "m" and "ch" maximal pair words

-Maximal Pair production with fishing game
-Sticker chart for effort

3. Describe any collaboration with the educational team and parent in selecting today's learning targets.

-Collaboration with the student's Pre-K teacher: The students are working on their writing skills in class. The Alphabeats program is commonly used in the classroom to assist the students with their writing skills. Therefore, it has been incorporated in today's speech session.
-Collaboration with the student's family: In our last IEP meeting, the student's parent let us know it can be difficult to reach her by phone, but I do send notes home (**show artifact**) and have regular contact with the student's older sister. Every other Friday, the student's older sister comes into the pre-K room for approx. 15 mins for reading buddies. Here, I can show her how to work with her brother at home to help him with his speech.

4. Describe the student's functional (i.e., "real-life", applicable to everyday experiences) communication needs across settings (e.g., speech room, regular education classroom, home, community, etc.).

The student's functional communication needs include:

-increasing his intelligibility of speech with a variety of communication partners (peers and adults) and in a variety of contexts (regular education classrooms and at home)

5. Describe how today's learning targets and lesson plan will relate/carryover to the student's functional communication needs described in question 4.

Today's learning targets will help to develop the student's phonological system so that he can be better understood in his classroom. Right now, his teacher understands only 50% of what says. His sister understands more of his speech, but acts as his interpreter often. As the student's phonological system develops, he will be able to produce the targeted sounds in phrases, sentences, stories and eventually in conversational speech, which will increase his overall intelligibility.

6. What will the observer see/hear to know you have planned for this session to be culturally relevant and personally meaningful to the student?

This session is personally meaningful to the student because:

-a game that he enjoys and is motivated by is being used to practice his targeted sounds.
-as the clinician reviews the targeted letters/sounds for today's session, she will address familiar vocabulary words that go along with each letter/sound (i.e. "m" is for mom, "c" is for car etc.).
-the student is in PreK, where letter names and sounds are being targeted daily in his reading curriculum

Additionally, I regularly I found the resource, African American English: An Overview by Anne Harper Charity that outlines specific linguistic characteristics of African American English (AAE) and provides examples of lexical, phonological, grammatical, and prosodic features of AAE and their implications for assessment. This resource is beneficial for working with this student because it increases my awareness of what is different about his speech and language versus disordered. In my first year as a speech pathologist in the Minneapolis Schools, I also attended a training on writing IEPs/evaluations and providing intervention with ELL students.
In ECSE, I attend a training about the Early Social Interaction Project through Florida State University.

After attending the training, I adapted a resource that was provided, during the training, to make it more relevant for Non-English speaking families that we work with. The adapted resource included visuals (**show artifact**).

7. What relevant resources and technology have you planned for today's session?

During this session, I will use the following materials:

-computer/ipad
-mirror
-visual aids/ visual schedule
-fishing game
-articulation flashcards
-magnet letters

8. Where in the session should I be looking for opportunities to assess student learning and how will you use the data you collect today to adapt instruction within today's session?

Opportunities to assess the students learning include:

-discrimination of "ch" vs. "m" sounds in words practice (percent accuracy of discrimination)
-making the "ch" sound in isolation while practicing writing and reviewing positioning of lips/tongue (percent accuracy of "ch" sound in isolation)
-practice of maximal pair targeted words during game using both imitative and spontaneous speech (percent accuracy of initial "ch" sound in words).

Treatment will occur with this sound until the student produced "ch" in the initial position of words with 90% accuracy over two consecutive sessions as outlined in the Gierut, 1989 article. Also, five contrasting "picturable" word pairs were be used in the game as outline by the Gierut, 1989 article

I will use the data to adapt instruction within today's session by varying the amount of time spent on each activity depending upon the data obtained. For example, if the student's percent accuracy of discriminating between words that start with the "m" sound and "ch"sound is low. I might make extra time to explicitly review where each sound is made in the mouth, and we might look in the mirror while saying the sounds as a visual prompt.

9. Is there anything else you'd like to share about the student and/or your lesson plans for today? This student can be shy, but I have told him that Ms. X is coming to visit on Thursday to help me become a better speech therapist. He is expecting you.

Domain 4 Questions (Professional Responsibilities, Due Process, & Case Management)

Dear SLP,

Please complete the form below with brief notes (feel free to use bullet points, short notes, etc.). In general, we'll spend approximately 2-3 minutes discussing each question during our pre-observation conference.

Thank you!

*The term "student" in the questions below may include an individual student, a group of students, or family/child. The term "service" may include assessment and evaluation, intervention and instruction, progress monitoring, and/or consultation and collaboration.

1. How have self-reflection and self-assessment of professional practices influenced your service to students? Please provide a specific example.
Self-reflection and self-assessment of my professional practices have influenced by service with students by helping me become aware of what I am doing well with my students and also helping me become aware of my areas of need. For example, I videotaped a part of a session with one of my students who stutters to collect/analyze a speech sample. As a reviewed the video, I not only assessed the student's speech, but I also assessed by own speech. During the video, I was observed to speak very quickly as I interacted with the student. I realized that if I am telling my student that he needs to practice using slow, smooth speech, I need to do the same. This self-assessment was a beneficial way for me to reflect on how I am providing services and gave me ideas of what I can do differently to better meet my student's needs.

Additionally, describe any tools and resources you have developed to enhance peer coaching, self-reflection, or self-assessment. Feel free to write "N/A" here:
N/A

2. How do you use professional feedback to improve your service to students? Please provide a specific example.
Professional feedback has improved my service to students by allowing me to gain ideas about what I can do differently and also by providing me with new intervention strategies/techniques. For example, after a recent peer observation with a colleague, I took the professional feedback she provided about ideas I could try to use to elicit more accurate vocalic /r/ sounds and implemented them into my next session with the student. I have also received professional feedback from peers about my report writing and goals/objectives. I have taken this feedback and have adapted some of my writing to include their suggestions which, in turn, directly improves the quality of service I provide to my students.

Additionally, describe any systems you have in place to gather regular feedback from colleagues, administrators, families, and students.
-regular team meeting to collaborate, answer/ask questions and to provide feedback about is going well and areas of need in regard to the implementation of services
-student surveys (distributed once per year) (**show artifact**)
-family surveys (distributed once per year)

3. How have you helped your colleagues use feedback to improve their services to students? Please provide a specific example.
I have helped my colleagues use feedback by providing examples of how use feedback. For example, during a recent team meeting, a colleague brought up a challenge she was experiencing with one of her young students on the Autism Spectrum. Florida State University has wonderful resources for supporting young students with Autism and I showed her where she could download and use those resources (**show artifact**).

4. How has professional development influenced your service to students? Please provide a specific example.
Professional development has influenced my service to students because it has provided me with knowledge of a variety of Evidenced Based Practices I can use with my students. Because I am more aware of these Evidence Based Practices, it has influenced my daily lesson plans and the service I provide to my students. For example, instead of using a traditional articulation approach with an articulation students, I will now use one of the Evidence Based Practices that were

presented during a professional development course for speech sound disorders (i.e. Minimal Pairs, Maximal Oppositions, Multiple Oppositions, etc.) depending on the severity of the disorder and the student's needs). Each of three EBP courses I attended this year have been extremely helpful.

Additionally, have you provided any professional learning opportunities for your colleagues? Feel free to write "N/A" here:
N/A

5. In what ways have you collaborated with families to improve your service to students? Please provide a specific example.
I have collaborated with families to improve service to students by:

-Discussing with family members both in person and over the phone about what they is going well in therapy, asking if they have seen any improvement in the student's communication skills and also asking about what they think are still areas of need for the student
-Providing services using Routines Based Intervention and talking to families about what routines they want to work on
-Sending home surveys for the families to fill out about their child's speech therapy services
-Sending home therapy materials that can be practiced at home

For example, after completing a project during speech therapy, a note will be sent home with the project/book stating; "Look at the book your son/daughter made! Please read this book with him/her to practice……." These is a way for me to communicate to the parents about what we are working on in speech therapy and it gives them the opportunity to practice the targeted skills with their child at home.

6. Walk me through the pre-referral process at your school/site.

-**show artifacts** (I will provide this walk through with visuals when we meet)

7. How do you ensure students receive an appropriate continuum of service in the Least Restrictive Environment? Please provide a specific example.
Model of service delivery is based on the student's specific needs.

-Student service should be decreased if the student is meeting his/her goals and no other areas of need are noted.
-Student service should be increased if the student is not making an expected amount of progress and if other factors (i.e. intervention strategies, group vs. one-on-one instruction) have been considered
-If a student has met his/her goals and no other areas of need are noted, the student should be dismissed from speech therapy services

Example: I am serving a fourth grader who has needs in the area of pragmatic language. He had met his current goals in the speech therapy room, so the LRE was changed to his general education classroom so he could generalize his learned skills with his peers and other adults in the classroom. His service time was also decreased from 60 minutes a week to 30 minutes a week, as he no longer

Additionally, describe any tools and resources have you developed for LRE. Feel free to write "N/A" here:
-**show artifacts** (I will show you some of the visuals used to support students in general education classrooms, when we meet)

8. Describe your systems (or, show artifacts) for collecting and summarizing data. Please describe any connections between these systems and EBP.

I print off a copy of the data collection sheet on EdPlan for each student. At the beginning of every session, I write the date and the targets for the session (i.e. initial /l/ in phrases). I then keep the data collection sheet close by and I record data throughout the session. I typically use "+" marks for accurate productions and "-" marks for inaccurate productions. At the end of the session, I totally up the "+" and "-" marks and calculate a percent accuracy for the targets. Periodically, I will plot the data on a graph to get a visual of the student's progress. My system aligns with EBP because it helps me evaluate the impact that the intervention I am using has on the student's progress. If I find that my intervention is not having a successful impact on my student, as shown by my data, I will revisit the EBP triangle and consider adjustments to better meet the student's needs.

9. How have you used data to make instructional decisions? Please provide a specific example for each of these two areas:

- **Assessment:** Weekly data I had collected for one of my students revealed that she had met her current language goals. Because the data suggested the student had met her goals, and there were no other areas of need noted, an evaluation was conducted in order to dismiss the student from services. I used both formal and informal assessment measures in that evaluation. (**show artifacts** for language sample probes)
- **Intervention:** After working on targeted vocalic /r/ sounds with a student for multiple sessions and not seeing progress in the data, as would be expected, I did some research and asked my colleagues for advice on different strategies/techniques I could use to help elicit more accurate sounds. The data I collected, helped me to realize I needed to change the intervention techniques I was using with this student.

Finally, for each area above, please describe any connections between your instructional decision and EBP.

EBP suggests you need to evaluate whether the clinical decision resulted in the anticipated impact for the student. In my "intervention" example, I stated that because the data did not suggest the anticipated impact of the student, I made the instructional decision to explore other intervention strategies.

EBP also suggests that for intervention and assessment the following steps need to be taken: a clinical question should be asked, you need to find evidence, you need to assess the evidence, you need to make a clinical decision and follow up with the decision. In the "assessment" example, I asked the clinical question of whether or not the student should be dismissed for speech and language service. I completed the assessment to gather evidence. I analyzed the results to assess the evidence and then, based on the evidence, made the clinical decision to dismiss the student from services. I followed up with this decision by presenting my evidence and reasoning to the IEP team.

Appendix D

Example of Completed Post-Observation Questions for the SLP in Appendix C

Full Observation: Post-observation Reflection Guide

Dear SLP,

Please complete the form below. We will discuss your responses at your post-observation and reflection conference.

Thank you!

1. What do you think went well in your observed speech/language session? What could make the speech/language session even stronger? Please provide specific examples and suggestions.

What went well…...

-I was able to complete all of my activities planned

-Many opportunities were provided for student to practice his "ch" sound

What could have been stronger….

-student was very easily distracted (an idea to help this would be to see the student more frequently for shorter sessions)

-less talking from the clinician (more wait time)

2. How did your assessment strategies provide data of student learning? To what extent did the student achieve the learning target(s) for the speech/language session? Provide data that you used to determine student learning.

-the assessment strategies provided many opportunities for me to keep track of accurate and inaccurate productions of the "ch" sound

-The student achieved learning targets by maintaining his accuracy production of the "ch" sound in words as compared to the previous session (23% accuracy in the initial position of words)

3. How will you use student performance in the speech/language session to plan future instruction? What will you do for student(s) who did not master the learning targets?

-As the accuracy of the student's productions in the initial position of words increases to 80% for two consecutive sessions, I will move on to a different target (another maximal pair or "ch" in the final position) as described by Gierut 1989

-If the student does not master the learning targets, I may try targeting different sounds using the maximal oppositions approach to see if the approach helps to develop the student's phonological system which in turn may generalize to the "ch" sound

Appendix E

Survey Provided to SLPs for their Evaluation of the Performance Evaluation Instrument

Dear Colleagues,
On the following pages, you will see the SLP evaluation rubric used in the 2014-2015 school year. Recall this rubric includes 27 elements organized into four domains:

1. Domain 1: Planning of Service
2. Domain 2: Climate of Service
3. Domain 3: Implementation of Service
4. Domain 4: Professional Responsibilities, Due Process Documentation, and Case Management

Having completed the SLP evaluation process this year, I am asking for your review of these elements and their scales (Requires Attention, Developing, Proficient, Exemplary). First, you will be asked to rate the extent to which you agree or disagree each ELEMENT represents an effective practice for SLPs. Then, you will be asked to rate the extent to which you agree or disagree the SCALES for this element accurately place an SLP along a continuum of performance.

This survey is ANONYMOUS and voluntary. The estimated time to complete this survey is 25 minutes. Thank you for your time.

<u>Page 1:</u>
ELEMENT i: Uses knowledge of Evidence-Based Practice (EBP) to plan the evaluation process

Requires Attention	Developing	Proficient	Exemplary
SLP does NOT demonstrate familiarity with EBP when selecting learning targets or planning implementation of service	*SLP demonstrates some familiarity with EBP when selecting learning targets and planning implementation of service*	*SLP applies EBP when selecting learning targets and planning implementation of service*	*SLP demonstrates extensive knowledge of EBP when selecting learning targets and planning implementation of service*
SLP cannot explain the EBP rationale for the selection of learning targets and planning of service implementation.	SLP struggles to explain the EBP rationale for the selection of learning targets and planning of service implementation, but is aware of EBP resources within or outside of the district.	1) SLP demonstrates the EBP process in the planning of service; **AND** 2) SLP seeks out and implements EBP training.	1) SLP models the EBP process in the planning of service (e.g., PD for SLPs and/or staff; explicit sharing of EBP references and processes in PLCs; parent/caregiver training explicitly related to EBP); **AND/OR** 2) SLP leads EBP training.

Please rate the extent to which you agree or disagree this ELEMENT represents an effective practice for SLPs:

- ❑ Strongly Agree
- ❑ Agree
- ❑ Disagree
- ❑ Strongly Disagree

Optional Comments:

Please rate the extent to which you agree or disagree the SCALES for this element accurately place an SLP along a continuum of performance:

- ❑ Strongly Agree
- ❑ Agree
- ❑ Disagree
- ❑ Strongly Disagree

Optional Comments:

Appendix F

Survey Provided to SLPs for their Evaluation of Evaluators

Dear Colleagues,
This survey will provide an opportunity to give feedback on your 2014-2015 observers.

This survey is ANONYMOUS and voluntary. The estimated time to complete this survey is 2-5 minutes.

Thank you for your time.

Please select your observer:

- ❑ X
- ❑ X
- ❑ X
- ❑ X
- ❑ X

1. My observer provided me with specific feedback:
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree
2. My observer provided clear evidence of my performance aligned to the SLP evaluation rubric:
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree
3. The time I spent with this observer was a valuable use of my time:
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree
4. I was able to have a reflective two-way dialogue with this observer:
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree
5. I felt respected as a professional with this observer during the observation process:
   - ❑ Strongly Agree
   - ❑ Agree
   - ❑ Disagree
   - ❑ Strongly Disagree

Optional comments:

Appendix G

Protocol for Coding Interrater Reliability Videos

Part I

1. Read Domain 1 preparation questions; take notes as needed
2. Watch video for Domain 1; take notes as needed
3. Score Domain 1. During scoring, video may be paused or replayed only under these circumstances:
    - Poor audio from student
    - Poor audio from SLP
    - Audio interference from outside source (e.g., overhead announcements)
    - Technical difficulties due to video camera

Part II (immediately following Part I)

1. Watch video for Domains 2 and 3; take notes as needed
2. Score Domains 2 and 3. During scoring, video may be paused or replayed only under these circumstances:
    - Poor audio from student
    - Poor audio from SLP
    - Audio interference from outside source (e.g., overhead announcements)
    - Technical difficulties due to video camera

Part III (immediately following Part II)

1. Read Domain 4 preparation questions; take notes as needed
2. Watch video for Domain 4; take notes as needed
3. Score Domain 4. During scoring, video may be paused or replayed only under these circumstances:
    - Poor audio from student
    - Poor audio from SLP
    - Audio interference from outside source (e.g., overhead announcements)
    - Technical difficulties due to video camera

Appendix H

Optional Survey Comments by Speech-Language Pathologists in the Evaluation of the 2014-2015 Performance Evaluation Instrument and their Evaluators

Table H1

Optional survey comments by speech-language pathologists (n = 47) in their evaluation of 2014-2015 instrument elements and indicators

	Comments	
Item	Element (*n* = 17)	Indicators (*n* = 56)
D1, E1: Uses knowledge of Evidence-Based Practice (EBP) to plan the evaluation process	It has raised awareness of why we do what we do even though it has been stressful recognizing that. It is very helpful to have articles and resources posted for researching/reading/application in the EBP areas each of us are implementing.	I don't know how often SLP's would have the chance to lead PD for EBP. This is not something that everyone is able to do in their buildings necessarily. Exemplary may not apply to all SLPs. What are non-subjective indicators (across observers/settings) that would determine that an SLP "cannot explain" vs. "struggles to explain"? I believe many SLPs are proficient in this area without necessarily seeking out additional training in EBP.
D1, E2: Designs coherent, sequential speech/language sessions that are aligned with student needs	(no comments)	I am not sure that "sequential" is the best word to describe how activities or lessons may build upon each other.
D1, E3: Designs speech/language sessions that place learning targets in functional contexts ("real-life", applicable to everyday experiences)	Functional and useful. Make this not too overly complicated just to sound complicated.	I think generalization across at least one environment would be proficient with two or more exemplary. Good element, but we do not always have access to a great deal of home information. How to address this? "Real world communication" - How is this defined?

		Some Artic students progress/exit faster with drill/practice focusing on sound targets, rather than necessarily "authentic." Would this result in possibility lower ratings for SLPs who serve students in those situations? Is the goal push-in for all? That might not apply to all disabilities/stages of intervention/parent wishes. I don't really see a clear difference between Proficient and Exemplary. Not all students are at a level of generalization in their skill progression. Evidence of collaboration is not always evident from an observation. If a student is not generalizing skills into other settings, because that might not be the goal on that day of observation, clinicians should not be rated as less proficient. It is good to think realistically where our students are headed and what skills we need to teach them to help them achieve their goals. The wording of this made it hard to plan for the speech session keeping in mind how I would show this. When I discussed with my observer, it became clearer to me what she was looking for. There has to be a different way to word Proficient & Exemplary to convey this.
D1, E4: Designs speech/language sessions that are culturally relevant and personally meaningful	(no comments)	We need to be cautious about how we define culturally relevant. In the Exemplary category: I think this is much more challenging to achieve than Exemplary in other elements. I'm not sure about the definition of "personally meaningful." Maybe should read "takes student's interests/skills into consideration?" Perhaps examples are needed for Exemplary?

D1, E5: Plans for relevant resources	One of the most functional, relevant, and realistic of the domains. This is in the SLPs control. Relevant resources is connected to workload.	This seems to indicate that them more resources you use, the better, which is not always the case--sometimes simpler is better. I think for most of these, the more specific the better! Liked this one! I don't think there is much of a difference between Proficient and Exemplary. Is the difference between Proficient and Exemplary? Perhaps differentiation or adaptation of relevant resources/technology in the classroom would be an Exemplary practice.
D1, E6: Plans for assessment strategies to monitor student progress	I think this is a very important aspect of service. However, I feel that realistically, I'm not sure it can be done with every student every session.	I'm not sure what would differentiate between Proficient and Exemplary? Is this referring to an action research project? Maybe other examples of Exemplary? Exemplary: I feel too much focus is being put using this phrase (EBP), or we need more support in understanding how to achieve Exemplary. Exemplary seems unattainable: Few clinicians have the time or opportunity to be planning action research around their many student groups. Unclear for Exemplary. How are "limited" vs. "sufficient" opportunities determined consistently across observers/settings/student disabilities? The Exemplary standard is not a realistic expectation the way is worded.

D2, E1: Establishes high expectations for student participation and explicitly sets up the environment so students understand the schedule and purpose of the service session	It is important for students to know and understand what they are learning, but this is hard for severe disabilities. It's a challenge for me. Measuring two different aspects: High expectations for student participation vs. purpose. Should be two different elements. Great element!	Too prescriptive. I think about errorless learning and how it meshes with "Let's try this..." I think the wording in the exemplary could be reconsidered. Using a statement such as "Today, we are going to... because..." is not always developmentally appropriate. More flexibility in the wording?
D2, E2: Establishes, maintains, and explicitly communicates room/space routines and procedures	I agree this is something we need to be aware of, but I don't think it needs to be rated on a rubric. This could be difficult for push-in service, where there may be limited opportunity for the SLP to control the room/space. If pull-out service, then I would Agree with this element.	(no comments)
D2, E3: Uses effective and constructive behavior management	(no comments)	Acknowledgement of expected student behavior as appropriate, to me, would be a standard, so therefore proficient. For the exemplary I would like to see something like "Positive feedback is the majority feedback given by the SLP. Behavior redirections result in a positive outcome." Hard to show if students did not experience behavior issues on the observation day. Are you then automatically proficient?
D2, E4: Builds positive relationships (i.e., rapport) with students	(no comments)	I disagree that "Makes no power-affirming remarks" is exemplary. That should be at proficient. This rating scale is too subjective. This is very important, but who is going to be negative during their observation? I would expect all to exemplary.
D3, E1: Explicitly communicates	I think this is important, but it may not be observed.	I like the specifics in this element.

speech/language learning objectives	This keeps us thinking about what we're doing each session & why.	I think we need to give training and more information about expectations (with examples) for this element. There seems to be some redundancy between this and an earlier element in Domain 2. Exemplary should not require written or visual information only. It should also include verbal information as well. Repeat from Domain 2? This element is too similar to an earlier one.
D3, E2: Provides learning activities and uses instructional strategies that are engaging and motivating	Functional and realistic element.	(no comments)
D3, E3: Provides opportunities to practice skills and demonstrate learning at a rigorous pace while ensuring student success	Useful	How is "pacing" measured? Is are implying pacing may be "too fast or slow" and can this be measured across observers/settings/students?
D3, E4: Uses techniques to promote cognitive engagement with learning targets	(no comments)	I need more clarification on wait time here. Wait time is critical. Exemplary is worded oddly I am trying to reconcile "errorless learning" with "above student's developmental level."
D3, E5: Uses assessment strategies to monitor student progress	(no comments)	"Adjusting instruction within an EBP framework" was unclear, until I received clarification from by observer. This could be more/less challenging depending on the area of communication you are working on (e.g. social communication vs. articulation). I would like more examples of how an SLP can "use data to

		make adjustments based on external EBP literature." How does this look on a week to week basis?
D3, E6: Gives explicit and timely feedback to improve student learning	(no comments)	(no comments)
D4, E1: Uses self-reflection and self-assessment to improve service to students	(no comments)	(no comments)
D4, E2: Uses feedback to improve service to students	(no comments)	A common challenge is reaching parents who may be highly mobile, but we just have to do our best and be creative as a school team.
D4, E3: Participates in relevant professional development	(no comments)	I believe it is important to attend PD regularly however I do not agree that is necessary to create and provide PD in order to be exemplary as a clinician. Are there options for Exemplary?
D4, E4: Collaborates with educational team	I appreciate the inclusion of this element. These things are hard to measure but so very important. The people we work with every day are the best judges of our performance. They see the big picture.	(no comments)
D4, E5: Collaborates with families	(no comments)	(no comments)
D4, E6: Applies knowledge of the pre-referral process in the prevention and identification of disabilities	This element may be less relevant in middle and high school but I agree that it is important.	Exemplary may not be meaningful for someone in birth-three. Might need examples for how to achieve for B-3.
D4, E7: Provides evaluations that are appropriate, accurate, and educationally-focused	(no comments)	I think Exemplary can be observed, as long as observers take into account different writing styles.

D4, E8: Proposes educational plans that are complete, educationally relevant, and measurable	(no comments)	(no comments)
D4, E9: Provides dynamic service delivery in the least restrictive environment	(no comments)	I believe this is very important, but the current observation rubric doesn't seem conducive to evaluating SLP push-in service.
D4, E10: Establishes data collection systems for all provided services	(no comments)	This is a little fuzzy for me. It would be helpful to have a clear description of the differences between levels here. I'm not sure how to measure "each" of the services as opposed to "some" of the services.
D4, E11: Makes data-based decisions (DBD) for all provided services	(no comments)	(no comments)

Table H2

Optional survey comments by speech-language pathologists (n = 73) in the evaluation of their 2014-2015 evaluators

Line Number	Regarding Evaluator	Comment
1	1	X did a wonderful job making this experience positive, professional and valuable!
2	1	X made me feel very comfortable. I was pretty nervous up until my observation day, but not as much after she arrived. X gave me specific ideas of what I could do to improve in certain areas rather than just telling me I had room for improvement. I learned quite a bit from our post-observation conversation.
3	1	X was very positive during this process. When X made comments and recommendations for constructive improvements they were respectfully given and helpful to me. She allowed enough time for me to ask questions that I had and ensure that I understood all aspects of the observation.
4	1	X was fabulous. X was very professional, but also approachable, positive and kind. I felt the process helped me to better understand how to improve in my growth areas and to value my strength areas. This would not have been true without having an observer like X.
5	1	Very positive in presentation. Collaborative. Easy to work with.
6	1	I really appreciated the time that X took to explain the ratings. I also enjoyed having a discussion that allowed me to brainstorm ways that I could motivate my students to use reading comprehension strategies more independently. X was very professional and objective throughout the whole process.
7	1	X was easy to talk to and not intimidating. The actual observation and feedback were not as stressful as it seemed it was going to be.
8	1	X was wonderful to work with on this!!
9	1	I think that the process was time consuming, but very beneficial. X was professional, positive, helpful and kind. X provided some helpful insights and useful tools for my future use.
10	1	X was a very fair, objective and positive observer. X provided me with specific feedback/ideas for areas for improvement (both self-identified and areas that came out of the observation process). Further, X was really wonderful about finding areas of strength that I didn't consider--applying the rubric criteria in ways I hadn't considered. Thanks X!
11	2	I thought X did a great job giving feedback, positive and negative, very constructively. I felt the observation process was a good tool to help me become more thoughtful and accountable in regard to my treatment planning and therapy. X was very supportive and reassuring.

12	2	X does a wonderful job with this process!
13	2	I saw a lot of care for me and for the task at hand. It was a very good opportunity to look at my "craft" with another professional in the field. The time spent with X to go over the observation was very helpful to provide a nice baseline for my performance. It's given me direction on where and how to grow moving forward.
14	2	I felt X took her time to make sure each part of the process was meaningful, clear, and helpful to me. I am very appreciative of her dedication to this observation process and to leading me through it with sound information, support, and positive encouragement!
15	2	I thought X did a great job of validating my skills and knowledge along with providing meaningful feedback to help me improve my practice. I appreciated her professionalism and how approachable X was.
16	2	Listened the majority of the time when I had a reason to doubt the score I received, and as a result some were changed. A little overly picky on the assessment and IEP, however, her feedback was constructive and helpful.
17	2	It was a positive experience but it did take a lot of time to complete.
18	2	Working with X was so helpful to my practice. The observation process was a walk in the park with her as my observer! Her expectations were clear, and X was flexible in scheduling.
19	3	X made the process both comfortable and helpful by being supportive, using active listening, answering my questions and giving evidence to support her ratings.
20	3	I felt that X was very supportive and made the experience more positive than I had anticipated.
21	3	I think X is a wonderful observer. X is so respectful, provides great feedback and is objective. X puts you at ease, which is so important.
22	3	X was great to work with. I really did learn from her and the process.
23	3	X was very flexible as my student was absent a couple of times when we planned the observation. X was understanding and we re-scheduled.
24	4	I did not feel threatened and was very comfortable with this observer.
25	4	I felt very comfortable and able to express any concerns with my observer. I felt the feedback X provided was helpful.
26	4	Observer was fair and helpful throughout process.
27	4	I actually enjoyed the process -- I didn't think I would. X made it comfortable.

28	4	I would have liked more time to finish the discussion; 30 mins is not enough for post conference. For next year, I would like more coaching opportunities added to the process. Coaching could be a combo of demonstration, co-teaching and observation. It would be nice to have a cadre of SLPs identified who could be available by appointment to do coaching on different areas, such as stuttering or AAC.
29	4	I found this experience to be uncomfortable. The observation felt unsupportive compared to other observation experiences I have had in the past and I got very little positive reinforcement to balance the areas that I needed to improve. Our post-interview felt one-sided. X is a great SLP, but her precise and strict observer style was not compatible with what I needed as an observee.
30	4	I felt respected. The atmosphere was collegial. Feedback was specific and instructional. Observer was professional.
31	5	Thanks for your valuable feedback.
32	5	Very helpful to me. X made the process fun.
33	5	My observation was a very positive and valuable experience.
34	5	Ideas that you shared with me have been so helpful in terms of improving my services to students and my collaboration with families. Thanks!
35	3	Very positive experience.
36	5	It was helpful to get clarity around due process for our students with unique cultural backgrounds.
37	5	The feedback was very constructive and I was able to apply it immediately.

Note: X = de-identified evaluator names.